BAYESIAN SEMIPARAMETRIC DENSITY DECONVOLUTION AND REGRESSION IN THE PRESENCE OF MEASUREMENT ERRORS

A Dissertation
by
ABHRA SARKAR

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Chair of Committee,	Bani K. Mallick
Co-Chair of Committee,	Raymond J. Carroll
Committee Members,	Anirban Bhattacharya
	Byung-Jun Yoon
Department Head,	Valen E. Johnson

August 2014

Major Subject: Statistics

ABSTRACT

Although the literature on measurement error problems is quite extensive, solutions to even the most fundamental measurement error problems like density deconvolution and regression with errors-in-covariates are available only under numerous simplifying and unrealistic assumptions. This dissertation demonstrates that Bayesian methods, by accommodating measurement errors through natural hierarchies, can provide a very powerful framework for solving these important measurement errors problems under more realistic scenarios. However, the very presence of measurement errors often renders techniques that are successful in measurement error free scenarios inefficient, numerically unstable, computationally challenging or intractable. Additionally, measurement error problems often have unique features that compound modeling and computational challenges.

In this dissertation, we develop novel Bayesian semiparametric approaches that cater to these unique challenges of measurement error problems and allow us to break free from many restrictive parametric assumptions of previously existing approaches. In this dissertation, we first consider the problem of univariate density deconvolution when replicated proxies are available for each unknown value of the variable of interest. Existing deconvolution methods often make restrictive and unrealistic assumptions about the density of interest and the distribution of measurement errors, e.g., normality and homoscedasticity and thus independence from the variable of interest. We relax these assumptions and develop robust and efficient deconvolution approaches based on Dirichlet process mixture models and mixtures of B-splines in the presence of conditionally heteroscedastic measurement errors. We then extend the methodology to nonlinear univariate regression with errors-in-covariates problems when the densities of the covariate, the regression errors and the measurement errors are all unknown, and the regression and the measurement errors are conditionally heteroscedastic. The final section of this dissertation is devoted to the development of flexible multivariate density deconvolution approaches. The methods available in the existing sparse literature all assume the measurement error density to be fully specified. In contrast, we develop multivariate deconvolution approaches for scenarios when the measurement error density is unknown but replicated proxies are available for each subject. We consider scenarios when the measurement errors are

distributed independently from the vector valued variable of interest as well as scenarios when they are conditionally heteroscedastic. To meet the significantly harder modeling and computational challenges of the multivariate problem, we exploit properties of finite mixture models, multivariate normal kernels, latent factor models and exchangeable priors in many novel ways.

We provide theoretical results showing the flexibility of the proposed models. In simulation experiments, the proposed semiparametric methods vastly outperform previously existing approaches. Our methods also significantly outperform theoretically more flexible possible nonparametric alternatives even when the true data generating process closely conformed to these alternatives. The methods automatically encompass a variety of simplified parametric scenarios as special cases and often outperform their competitors even in those special scenarios for which the competitors were specifically designed. We illustrate practical usefulness of the proposed methodology by successfully applying the methods to problems in nutritional epidemiology. The methods can be readily adapted and applied to similar problems from other areas of applied research. The methods also provide the foundation for many interesting extensions and analyses.

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest gratitude towards my advisors Dr. Bani K. Mallick and Dr. Raymond J. Carroll. Throughout, they have been extremely generous, caring and supportive. They kept faith in my abilities, gave me enough freedom and opportunities so that I could pursue my own ideas and interests, and put me back in track whenever I digressed a bit too far. I am also extremely grateful to them for continually providing me with financial support through their research grants.

I am extremely fortunate to have had the opportunity to work with Dr. Carroll. He radiates so much positive energy and enthusiasm that no matter how big a problem I would be stuck in, I could always find the courage and motivation to work my way out after I discussed it with him. To emulate his limitless passion towards academics may never be possible, but I consider my life a success to try so continually.

I am also immensely grateful to Dr. Mallick. I cannot thank him enough for helping me start afresh in a completely new environment, for introducing and promoting me to great opportunities, and for being always ears to my problems, however small or insignificant they were.

I am especially fortunate to have crossed paths with Avishek-da (Dr. Avishek Chakraboty) and Debdeep-da (Dr. Debdeep Pati). They have played major roles in my academic, professional and personal growth and I cannot find enough words to thank them. I have learnt a lot collaborating with Debdeep-da on some of my dissertation chapters and sincerely acknowledge his patience in mentoring me. The bond and the friendship I share with them are some of my most proud possessions.

I sincerely thank Dr. Byung-Jun Yoon for his help and support and for teaching a wonderful course on probabilistic graphical models. My other committee member Anirban-da (Dr. Anirban Bhattacharya) has become more of a friend since he joined TAMU last year. I thank him for his guidance on academic matters and also for the many beautiful moments we shared together over lunch, dinner or tea.

I am grateful to Dr. David Dahl for introducing me to the fascinating world of Bayesian nonparametrics. I thank Dr. Soumendra Nath Lahiri, Dr. Mohsen Pourah-

madi, Dr. Samiran Sinha, Dr. Suhasini Subbarao, Dr. Yanyuan Ma, Dr. Jeffrey Hart, Dr. Valen Johnson, Dr. Faming Liang and Dr. John Staudenmayer for their support and encouragement. Special thanks to Dr. Mike Longnecker for mentoring me when I taught courses or worked as a TA. I would like to take this opportunity to thank the administrative stuff of our department, in particular Maryline, Sandra, Athena, Joyce and Elaine for always being so nice and helpful. I am obliged to the National Science Foundation (NSF), the American Statistical Association (ASA), the King Abdullah University of Science and Technology (KAUST) for financing my graduate education and conference travel. I acknowledge financial assistance offered by the department in the form of teaching assistantships and travel grants. I also acknowledge the excellent computing infrastructure at TAMU including that of our department and the Brazos HPC supercomputing cluster that contributed to my research.

There are so many special people who have touched my life at TAMU. I am extremely grateful to Krishna-da (Krishna Ganesula) and Anirban-da (Dr. Anirban Mondal) for helping me grow as a human being. When I first arrived here, to a new culture, among new people, I may have come across as a reclusive personality. I owe a lot to Krishna-da for sticking with me and helping me settle in. I specially thank Robyn (Dr. Robyn Ball, the best officemate anyone could ask for) and Yanqing (Yanqing Wang) for their support during times of hardships. I would like to separately mention Paritosh-da (Paritosh Pande), Mrinmoy-da (Dr. Mrinmoy Chakraborty) and Shyamalendu with whom I shared wonderful memories. I thank Swarup-da, Arindam-da, Sulagna-di, Parijat-da, Debu-da, Manasi-di, Papiya-di, Kashyap-da, Paromita, Debkumar-da, Dhruba-di, Sabyasachi-da, Aditi-di, Sharmila-di, Suprateek-da, Debanjan-da, Adrija-da, Soma-da, Rajesh-da, Soutir-da, Ayan-da, Anindya-da, Deep-da, Shahina, Soutrik, Antik, Kun Xu, Yichen Cheng, Raniye Sun, Cassy Wang and Rubin Wei for making me feel home away from home. They all conspired successfully to make life in the States enjoyable, happening, and provided me with sweet memories to cherish.

I would also like to thank my teachers and friends from India, without whose support and encouragement I could not have come this far. Dr. Gourangdeb Chattopadhyay and Ramendra Maharaj deserve special mention, for always being there for me whenever I needed their support and guidance. Bappa, Sayantan, Puspendu, Moumita, Jyotirmoy, Rivu and Kaushik have been wonderful and supportive friends

throughout my journey.

I thank my parents Sukdeb and Kavery Sarkar, my sweetest Rakhi-didi and my wonderful brother Arghya for their unconditional love, support and sacrifices. Arghya, although six years younger than me, has often simplified things for me with a fresh approach. I thank my wife and best friend Rimli for being so loving and caring, for bringing balance and stability in my life, for being by my side through thick and thin, for sacrificing so much for me. Finally, I would like to remember my recently deceased grandparents Ratikanta Sarkar and Bijali Sarkar. They showered me with all their love and affection and gave me a wonderful childhood. They had always wanted me to become a good scholar and I so wish that they could live to see this day.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Development of flexible and efficient Bayesian semiparametric methodology for fundamental measurement error problems is the primary contribution of this dissertation. Although the literature on measurement error problems is extensive, solutions to even the most fundamental measurement error problems like density deconvolution and regression with errors-in-predictors were available only under numerous simplifying and unrealistic assumptions. We demonstrate in this dissertation that Bayesian methods, by accommodating measurement errors through natural hierarchies, can provide a very powerful framework for solving these important measurement errors problems under less restricted and more realistic scenarios. However, the very presence of measurement errors often renders techniques that are successful in measurement error free scenarios inefficient, numerically unstable, computationally challenging or intractable. Additionally, measurement error problems often have unique features that compound modeling and computational challenges. For instance, we show that nonparametric techniques designed for modeling conditionally varying regression errors that allow all aspects of the error distribution to vary flexibly with the conditioning variable either become numerically highly unstable in the absence of precise measurements on the conditioning variable, or are not at all relevant due to the characteristic differences between regression and measurement errors. In similar vein, Gaussian process based nonparametric regression techniques, though immensely popular and successful in measurement error free scenarios, are not suitable for regression with errors-in-covariates problems.

In this dissertation, we address these unique challenges by developing Bayesian semiparametric approaches that make fairly minimal and highly plausible structural assumptions on the measurement errors and take into account their prominent and unique features. Compared to possible nonparametric alternatives, the proposed semiparametric approaches are numerically much more stable and hence are better suited for measurement error problems. In theory they are less flexible than possible nonparametric alternatives, but they still allow us to break free from many restrictive parametric assumptions of existing approaches. More importantly, the methods are highly robust to departures from the assumed structural assumptions. In simulation experiments, the proposed semiparametric methods significantly out-

perform theoretically more flexible nonparametric alternatives even when the true data generating process closely conformed to those alternatives. The methods also automatically accommodate a variety of simplified parametric scenarios as special cases and often outperform their competitors even in those simplified scenarios for which the competitors are specifically designed.

We present below an overview of each major section of this dissertation. More detailed outlines and more detailed review of relevant literature and their limitations are provided at the beginning of each major section.

In Section 2, we develop univariate density deconvolution approaches when replicated proxies are available for each unknown value of the variable of interest and the variability of the measurement errors depends on the associated unobserved value of the variable of interest through an unknown relationship. We assume that the measurement errors could be factored into a variance function that explains conditional heteroscedasticity and scaled errors that are independent of the variable of interest. We model the density of interest using flexible location-scale mixtures of normal kernels induced by a Dirichlet process. The density of the scaled errors is also modeled by Dirichlet process induced mixtures, where the component kernels themselves were two component mixtures of normals with their means restricted at zero. The use of moment restricted two-component mixtures of normals as components restricts the mixture density to have zero mean but gives it flexibility to model other unconstrained aspects of the distribution of scaled errors. The variance function is modeled by flexible smooth mixtures of B-splines. While being extremely flexible in its capacity to adapt to asymmetry, heavy tails and multimodality, the proposed deconvolution approach can also accommodate commonly assumed parametric models as special cases.

In Section 3, we extend the methodology to a univariate errors-in-variables regression problem. The latent covariate and conditionally heteroscedastic regression and measurement errors are modeled using Dirichlet process induced mixtures as in Section 2. The regression function is modeled using flexible mixtures of B-splines. The proposed model can accommodate normally distributed covariate and homoscedastic or ordinary heteroscedastic and/or normally distributed regression and measurement errors as special cases.

Section 4 of this dissertation is devoted to the development of flexible multivariate density deconvolution approaches. To the best of our knowledge, the few

methods available in the extremely sparse literature on the problem of multivariate deconvolution all consider very specific and restrictive scenarios where the measurement errors are assumed to be distributed independently of the vector of interest according to some fully specified parametric probability law. We propose robust Bayesian semiparametric multivariate deconvolution approaches for scenarios when the measurement error density is not known but replicated proxies are available for each unobserved value of the random vector. We allow the measurement errors to be distributed independently of the vector of interest as well as to be conditionally heteroscedastic. Straightforward extensions of the univariate deconvolution approaches developed in Section 2 fail to meet the significantly harder modeling and computational challenges of the multivariate problem. Instead of using infinite mixtures induced by Dirichlet processes, for the multivariate problem we use finite mixtures of multivariate normal kernels with exchangeable priors on the mixture probabilities that enable us to significantly reduce computational complexity while retaining essentially the same flexibility of infinite dimensional models. Exploiting the exchangeability of the symmetric Dirichlet prior and basic properties of multivariate normal kernels, we propose a novel mechanism to enforce the mean zero restriction on the density of the measurement errors. From a computational point of view, the mechanism is much easier to implement than the technique adopted in Section 2 and hence more suitable for multivariate problems. We show that, due to their dense parametrization, inverse Wishart priors on the component specific covariance matrices are not suitable for deconvolution problems, particularly when the measurement errors are conditionally heteroscedastic and the likelihood function becomes fairly complicated. Using factor-analytic representations of the component specific covariance matrices with sparsity inducing shrinkage priors on the factor loading matrices, we build models that are flexible yet parsimonious and numerically stable and lead to significant improvements in performance. To meet additional computational challenges posed by conditionally heteroscedastic multivariate measurement errors, we design a novel two-stage procedure to estimate the parameters. We first estimate the variance functions using reparametrized versions of the corresponding univariate submodels. We then estimate the remaining model parameters in the second stage, plugging in the estimates obtained in the first stage and keeping them fixed.

Theoretical results showing the flexibility of the proposed methods are provided. We evaluate the performance of the proposed methods through extensive simulation

experiments and show that the deconvolution and regression methodology developed in Section 2 and Section 3 vastly outperform their competitors that are often considered to be the current gold standards. The proposed methods also often outperform their competitors in simplified parametric scenarios for which the competitors were specifically designed. The multivariate deconvolution methods developed in Section 4 have essentially no competitors to compare their performance with, but they show good empirical performance in the simulation experiments.

While the methods can be readily adapted and applied to other areas of applied research, in this dissertation we focus on problems in nutritional epidemiology, where the assessment of long term dietary habits and related questions are of extreme importance. In the absence of sophisticated methods, nutritionists often apply transformation techniques to make the data conform more closely to normality and homoscedasticity, then analyze data on the transformed scale using existing parametric methods, and finally transform the results back to the original scale. In this dissertation we demonstrate that it is seldom possible to transform data to additivity, homoscedasticity and normality. Methods based on multiple non-linear transformations and approximations thus often result in the masking of important features of the data. The methods developed in this dissertation, on the other hand, are extremely flexible and can operate on both observed and transformed scales. The univariate and multivariate deconvolution methods developed in Section 2 and 4 are applied to estimate consumption patterns of different dietary components from contaminated 24 hour recalls and are found to be extremely useful for uncovering interesting features in long term consumption patterns of different dietary components. Similarly, the regression techniques developed in Section 3 are successfully employed to understand the regression relationship between the intakes reported by respondents in food frequency questionnaires and their true dietary intakes.

2. DENSITY DECONVOLUTION IN THE PRESENCE OF CONDITIONALLY HETEROSCEDASTIC MEASUREMENT ERRORS *

2.1 Introduction

Many problems of practical importance require estimation of the unknown density of a random variable. The variable, however, may not be observed precisely, observations being subject to measurement errors. Under the assumption of additive measurement errors, the observations are generated from a convolution of the density of interest and the density of the measurement errors. The problem of estimating the density of interest from available contaminated measurements then becomes a problem of deconvolution of densities.

In this section, we propose novel Bayesian semiparametric approaches for robust estimation of the density of interest when the measurement error density is not known and the variability of the measurement errors depends on the associated unobserved value of the variable of interest through an unknown relationship. We assume that replicated proxies are available for each unobserved value of the variable of interest. The proposed methodology is fundamentally different from existing deconvolution methods, relaxes many restrictive assumptions of existing approaches by allowing both the density of interest and the distribution of measurement errors deviate from standard parametric laws and also by accommodating conditional heteroscedasticity, and significantly outperforms previous methodology.

Most of the early literature on density deconvolution considers scenarios when a single contaminated measurement is available for each subject. To make the density of interest identifiable from the observed data, these methods typically assume that the measurement errors are independently and identically distributed according to some known probability law (often normal). Kernel deconvoluting approaches have been studied by Stefanski and Carroll (1990), Carroll and Hall (1988), Liu and Taylor (1989), Devroye (1989), Fan (1991a, 1991b, 1992) and Hesse (1999) among others. The nonparametric maximum likelihood (NPML) approach of Kiefer and

*Part of this section is from "Bayesian Semiparametric Density Deconvolution in the Presence of Conditionally Heteroscedastic Measurement Errors" by Sarkar, A., Mallick, B. K., Staudenmayer, J., Pati, D. and Carroll, R. J. (2014). Forthcoming in *Journal of Computational and Graphical Statistics*. DOI:10.1080/10618600.2014.899237. Copyright 2014 by Carroll, R. J. Reprinted by permission of Taylor & Francis LLC (http://www.tandfonline.com).

Wolfowitz (1956) estimates the distribution of interest by maximizing the likelihood of the observed data under a known measurement error distribution. The resulting estimates are, however, discrete (Liard, 1978) and are, therefore, unsuitable for many practical applications. See also Lindsay (1983a, 1983b), Böhning, et al. (1998), Pilla and Lindsay (2001) and Section 10.2.3 of Buonaccorsi (2010). Other deconvolution approaches that do not make rigid assumptions about the density of interest but assume a known measurement error distribution include the wavelet based methods of Pensky and Vidakovic (1999) and Fan and Koo (2002).

Of course, in reality the distribution of measurement errors is rarely known, and the assumption of constant variance measurement errors is also often unrealistic. Misspecification of the distribution of measurement errors may lead to biased and inefficient estimates of the density of interest. The focus of recent deconvolution literature has, therefore, been on robust deconvolution methods under less restrictive assumptions on the error distribution. These methods typically require the availability of replicated proxies for each unknown value of the variable of interest.

The problem of deconvolution when errors are homoscedastic with an unknown density has been addressed by a few authors. Li and Vuong (1998) showed that the characteristic functions of both the variable of interest and the measurement errors can be estimated from the empirical characteristic function of replicated contaminated data. Inverse Fourier transforms of these estimated characteristic functions then provide nonparametric estimators of the corresponding densities. Carroll and Hall (2004) considered the problem of estimating a low-order approximation of the density of interest, rather than the density itself. Their method required estimates of low-order moments of the measurement error distribution, which can be estimated from replicated proxies. See also Diggle and Hall (1993), Neumann (1997) and Delaigle, et al. (2008).

More recently the problem of deconvolution with measurement errors that are heteroscedastic and have a known distribution has also received some attention. Delaigle and Meister (2008) and McIntyre and Stefanski (2011) proposed generalizations of kernel density estimators. See also Wang and Wang (2010). The NPML approach can also be generalized to allow different variances for different subjects (DerSimonian, 1986).

All the above-mentioned deconvolution approaches assume that the measurement errors are independent of the variable of interest. Staudenmayer, et al. (2008)

relaxed this assumption and considered the problem of density deconvolution in presence of conditionally heteroscedastic measurement errors. The density of interest was modeled by a penalized positive mixture of normalized quadratic B-splines. Measurement errors were assumed to be normally distributed but the measurement error variance was modeled as a function of the associated unknown value of the variable of interest using a penalized positive mixture of quadratic B-splines.

The focus of this section of the dissertation is also on deconvolution in the presence of conditionally heteroscedastic measurement errors, but the proposed Bayesian semiparametric methods are vastly different from the approach of Staudenmayer, et al. (2008), as well as from other existing methods. The density of interest is modeled by a flexible location-scale mixture of normals induced by a Dirichlet process (Ferguson, 1973). For modeling conditionally heteroscedastic measurement errors, it is assumed that the measurement errors can be factored into 'scaled errors' that are independent of the variable of interest and have zero mean and unit variance, and a 'variance function' component that explains the conditional heteroscedasticity. This multiplicative structural assumption on the measurement errors was implicit in Staudenmayer, et al. (2008), where the scaled errors were assumed to come from a standard normal distribution. Our approach is based on a more flexible representation of the scaled errors. The density of the scaled measurement errors is modeled using an infinite mixture model induced by a Dirichlet process, each component of the mixture being itself a two-component normal mixture with its mean restricted at zero. The use of moment restricted two-component mixture of normals as components restricts the mixture density to have zero mean but gives it flexibility to model other unconstrained aspects of the distribution of scaled errors. This deconvolution approach, therefore, uses flexible Dirichlet process mixture models twice, first to model the density of interest and second to model the density of the scaled errors, freeing them both from restrictive parametric assumptions, while at the same time accommodating conditional heteroscedasticity through the variance function. Although in the deconvolution literature there are instances of using finite mixtures of normalized spline densities to model (only) the density of interest, infinite mixture models, to the best of our knowledge, have not been used.

In the Bayesian paradigm, theoretical properties of nonparametric heteroscedastic error models have recently been studied by Pati and Dunson (2013) and Pelenis (2014) in a regression context. However, in a deconvolution problem, it is not clear

whether the variance function can be suitably identified and efficiently estimated under their modeling assumptions. In the deconvolution context, it is, therefore, of importance to know under what conditions a simpler and computationally efficient heteroscedastic formulation will be flexible. In our method, efficiency is achieved by taking a semiparametric route as described above. Flexibility of the proposed formulation in modeling the implied marginal, conditional and joint densities of interest is theoretically investigated assuming a similar multiplicative structure for the truth. Empirical evidence is also provided to show that in deconvolution problems the proposed semiparametric approach would often be more efficient than possible nonparametric alternatives, even when the truth departs from the assumed multiplicative structure.

The section is organized as follows. Section 3.2 details the models. Section 2.3 discusses some model diagnostic tools. Section 2.4 discusses the choice of hyperparameters. Section 2.5 describes MCMC methods to sample from the posterior. Section 3.3 presents extensive simulation studies comparing the proposed semiparametric methods with the method of Staudenmayer, et al. (2008) and a possible nonparametric alternative. Section 3.4 presents an application of the proposed methodology in estimation of the distributions of daily dietary intakes from contaminated 24 hour recalls in a nutritional epidemiologic study. Section 3.5 contains concluding remarks. Model identifiability and results of additional simulation experiments are discussed in Appendix A. Appendix D provides a theoretical study of the flexibility of the proposed models.

2.2 Density Deconvolution Models

The goal is to estimate the unknown density of a random variable X. There are $i = 1, 2, \ldots, n$ subjects. Precise measurements of X are not available. Instead, for $j = 1, 2, \ldots, m_i$, replicated proxies W_{ij} contaminated with heteroscedastic measurement errors U_{ij} are available for each subject. The replicates are assumed to be generated by the model

$$W_{ij} = X_i + U_{ij}, \tag{2.1}$$

$$U_{ij} = v^{1/2}(X_i)\, \epsilon_{ij}, \tag{2.2}$$

where X_i is the unobserved true value of X; ϵ_{ij} are independently and identically distributed with zero mean and unit variance and are independent of the X_i, and v

is an unknown smooth variance function. Identifiability of model (1)-(2) is discussed in A.1, where we show that 3 replicates more than suffices. Some simple diagnostic tools that may be employed in practical applications to assess the validity of the structural assumption (2) on the measurement errors are discussed in Section 2.3.

The density of X is denoted by f_X. The density of ϵ_{ij} is denoted by f_ϵ. The implied conditional distributions of W_{ij} and U_{ij}, given X_i, is denoted by the generic notation $f_{W|X}$ and $f_{U|X}$, respectively. The marginal density of W_{ij} is denoted by f_W.

Model (2), along with the moment restrictions imposed on the scaled errors ϵ_{ij}, implies that the conditional heteroscedasticity of the measurement errors is explained completely through the variance function v, while other features of $f_{U|X}$ are derived from f_ϵ. In a Bayesian hierarchical framework, model (1)-(2) reduces the problem of deconvolution to three separate problems: (a) modeling the density of interest f_X; (b) modeling the variance function v, and (c) modeling the density of the scaled errors f_ϵ.

2.2.1 Modeling the Distribution of X

We use Dirichlet process mixture models (DPMMs) (Ferguson, 1973, Escobar and West, 1995) for modeling f_X. For modeling a density f, a DPMM with concentration parameter α, base measure P_0, and mixture components coming from a parametric family $\{f_c(\cdot \mid \boldsymbol{\phi}) : \boldsymbol{\phi} \sim P_0\}$, can be specified as

$$f(\cdot) = \textstyle\sum_{k=1}^{\infty} \pi_k \; f_c(\cdot \mid \boldsymbol{\phi}_k), \quad \boldsymbol{\phi}_k \sim P_0, \quad \pi_k = s_k \prod_{j=1}^{k-1}(1 - s_j), \quad s_k \sim \text{Beta}(1, \alpha).$$

In the literature, this construction of random mixture weights $\{\pi_k\}_{k=1}^{\infty}$ (Sethuraman, 1994), is often represented as $\pi \sim \text{Stick}(\alpha)$. DPMMs are, therefore, mixture models with a potentially infinite number of mixture components or 'clusters'. For a given data set of finite size, however, the number of active clusters exhibited by the data is finite and can be inferred from the data.

Choice of the parametric family $\{f_c(\cdot \mid \boldsymbol{\phi}) : \boldsymbol{\phi} \sim P_0\}$ is important. Mixtures of normal kernels are, in particular, very popular for their flexibility and computational tractability (Escobar and West, 1995; West, et al. 1994). We also specify f_X as a mixture of normal kernels, with a conjugate normal-inverse-gamma (NIG) prior on

the location and scale parameters

$$f_X(X) = \sum_{k=1}^{\infty} \pi_k \text{ Normal}(X \mid \mu_k, \sigma_k^2), \tag{2.3}$$

$$\pi \sim \text{Stick}(\alpha_X), \qquad (\mu_k, \sigma_k^2) \sim \text{NIG}(\mu_0, \sigma_0^2/\nu_0, \gamma_0, \sigma_0^2). \tag{2.4}$$

Here $\text{Normal}(\cdot \mid \mu, \sigma^2)$ denotes a normal distribution with mean μ and standard deviation σ. In what follows, the generic notation p_0 will sometimes be used for specifying priors and hyper-priors.

2.2.2 Modeling the Variance Function

Examples of modeling log-transformed variance functions using flexible mixtures of splines are abundant in the literature. Yau and Kohn (2003), for example, modeled $\log\{v(X)\}$ using flexible mixtures of polynomial and thin-plate splines. Liu, et al. (2007) proposed a penalized mixture of smoothing splines, whereas Chan, et al. (2006) considered mixtures of locally adaptive radial basis functions.

We model the variance function as a positive mixture of B-spline basis functions with smoothness inducing priors on the coefficients. For a given positive integer K, partition an interval $[A, B]$ of interest into K subintervals using knot points $t_1 = \cdots = t_{q+1} = A < t_{q+2} < t_{q+3} < \cdots < t_{q+K} < t_{q+K+1} = \cdots = t_{2q+K+1} = B$. For $j = (q+1), \ldots, (q+K)$, define $\Delta_j = (t_{j+1} - t_j)$ and $\Delta_{\max} = \max_j \Delta_j$. It is assumed that $\Delta_{\max} \to 0$ as $K \to \infty$. Using these knot points, $(q+K) = J$ B-spline bases of degree q, denoted by $\mathbf{B}_{q,J} = \{b_{q,1}, b_{q,2}, \ldots, b_{q,J}\}$, can be defined through the recursion relation given on page 90 of de Boor (2000), see Figure 2.1. A flexible model for the variance function is

$$v(X) = \sum_{j=1}^{J} b_{q,j}(X) \exp(\xi_j) = \mathbf{B}_{q,J}(X) \exp(\boldsymbol{\xi}), \tag{2.5}$$

$$p_0(\boldsymbol{\xi} \mid J, \sigma_\xi^2) \propto \exp\{-\boldsymbol{\xi}^{\mathrm{T}} P \boldsymbol{\xi}/(2\sigma_\xi^2)\}, \tag{2.6}$$

$$p_0(\sigma_\xi^2) = \text{IG}(a_\xi, b_\xi), \quad K \sim p_0(K). \tag{2.7}$$

Here $\boldsymbol{\xi} = \{\xi_1, \xi_2, \ldots, \xi_J\}^{\mathrm{T}}$; $\exp(\boldsymbol{\xi}) = \{\exp(\xi_1), \exp(\xi_2), \ldots, \exp(\xi_J)\}^{\mathrm{T}}$, and $\text{IG}(a, b)$ denotes an inverse-Gamma distribution with shape parameter a and scale parameter b. We choose $P = D^{\mathrm{T}} D$, where D is a $J \times (J+2)$ matrix such that $D\boldsymbol{\xi}$ computes the second differences in $\boldsymbol{\xi}$. The prior $p_0(\boldsymbol{\xi} \mid \sigma_\xi^2)$ induces smoothness in the coefficients because it penalizes $\sum_{j=1}^{J} (\Delta^2 \xi_j)^2 = \boldsymbol{\xi}^{\mathrm{T}} P \boldsymbol{\xi}$, the sum of squares of the second order differences in $\boldsymbol{\xi}$ (Eilers and Marx, 1996). The variance parameter σ_ξ^2 plays the role of

smoothing parameter - the smaller the value of σ_ξ^2, the stronger the penalty and the smoother the variance function. The inverse-Gamma hyper-prior on σ_ξ^2 allows the data to have strong influence on the posterior smoothness and makes the approach data adaptive. The prior $p_0(K)$ assigns positive probability to all $K \in \mathbb{N}$, the set of all positive integers.

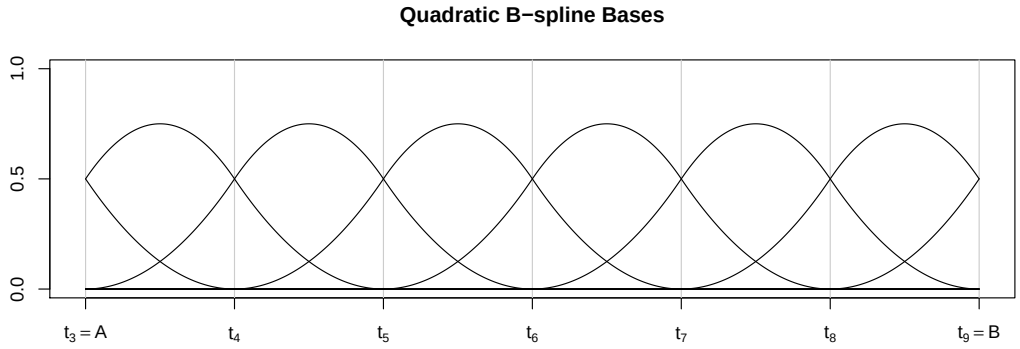

Figure 2.1: Plot of 9 quadratic ($q = 2$) B-splines on $[A, B]$ defined using 11 knot points that divide $[A, B]$ into $K = 6$ equal subintervals.

2.2.3 Modeling the Distribution of the Scaled Errors

Three different approaches of modeling the density of the scaled errors f_ϵ are considered here, successively relaxing the model assumptions as we progress.

2.2.3.1 Model-I: Normal Distribution

We first consider the case where the scaled errors are assumed to follow a standard normal distribution

$$f_\epsilon(\epsilon) = \text{Normal}(\epsilon \mid 0, 1). \tag{2.8}$$

This implies that the conditional density of measurement errors is given by $f_{U|X}(U \mid X) = \text{Normal}\{U \mid 0, v(X)\}$. Such an assumption was made by Staudenmayer, et al. (2008).

2.2.3.2 Model-II: Skew-Normal Distribution

The strong parametric assumption of normality of measurement errors may be restrictive and inappropriate for many practical applications. As a first step towards modeling departures from normality, we propose a novel use of skew-normal distributions (Azzalini, 1985) to model the distribution of scaled errors. A random variable Z following a skew-normal distribution with location ξ, scale ω and shape parameter λ has the density $f(Z) = (2/\omega)\phi\{(Z-\xi)/\omega\}\Phi\{\lambda(Z-\xi)/\omega\}$. Here ϕ and Φ denote the probability density function and cumulative density function of a standard normal distribution, respectively. Positive and negative values of λ result in right and left skewed distributions, respectively. The Normal$(\cdot \mid \mu, \sigma^2)$ distribution is obtained as special cases with $\lambda = 0$, whereas the folded normal or half-normal distributions are obtained as limiting cases with $\lambda \to \pm\infty$, see Figure 2.2. With $\delta = \lambda/(1+\lambda^2)^{1/2}$, the mean and the variance of this density are given by $\mu = \xi + \omega\delta(2/\pi)^{1/2}$ and $\sigma^2 = \omega^2(1 - 2\delta^2/\pi)$, respectively. Although the above parametrization is more constructive and intuitive in revealing the relationship with the normal family, we consider a different parametrization in terms of μ, σ^2 and λ, denoted by SN$(\cdot \mid \mu, \sigma^2, \lambda)$, that is more useful for specifying distributions with moment constraints, namely $f(Z) = (2\zeta_2/\sigma)\phi\{\zeta_1 + \zeta_2(Z-\mu)/\sigma\}\Phi[\lambda\{\zeta_1 + \zeta_2(Z-\mu)/\sigma\}]$, where $\zeta_1 = \delta(2/\pi)^{1/2}$ and $\zeta_2 = (1 - 2\delta^2/\pi)^{1/2}$. For specifying the distribution of the scaled errors we now let

$$f_\epsilon(\epsilon) = \text{SN}(\epsilon \mid 0, 1, \lambda), \tag{2.9}$$

$$p_0(\lambda) = \text{Normal}(\lambda \mid \mu_{0\lambda}, \sigma^2_{0\lambda}). \tag{2.10}$$

The implied conditionally heteroscedastic, unimodal and possibly asymmetric distribution for the measurement errors is given by $f_{U|X}(U \mid X) = \text{SN}\{U \mid 0, v(X), \lambda\}$.

2.2.3.3 Model-III: Infinite Mixture Models

While skew-normal distributions can capture moderate skewness, they are still quite limited in their capacity to model more severe departures from normality. They can not, for example, model multimodality or heavy tails. In the context of regression analysis, moment constrained infinite mixture models have recently been used by Pelenis (2014) (see also the references therein) for flexible modeling of error distributions that can capture multimodality and heavy tails. They considered the mixture

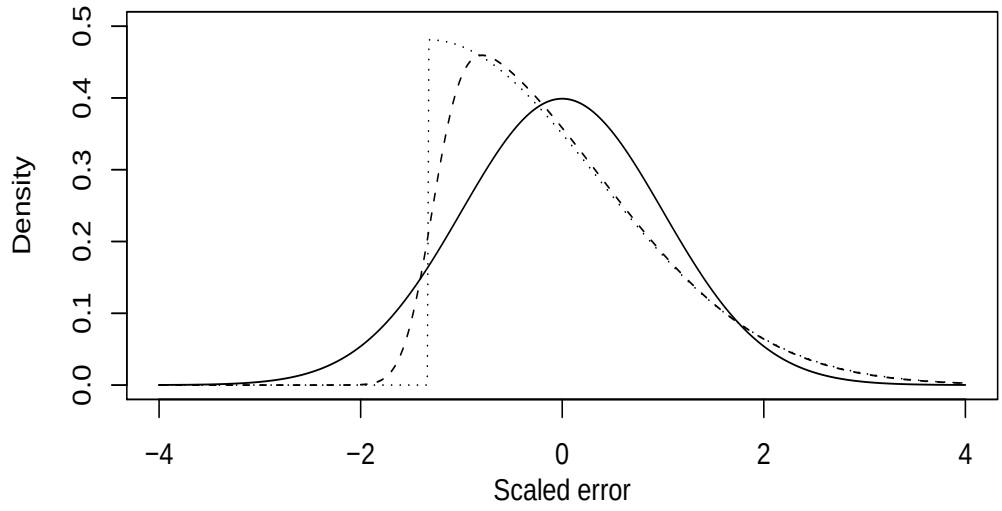

Figure 2.2: Skew-normal densities with mean=0, variance=1 and varying skewness parameter λ. The solid line is the density of $\mathrm{SN}(\cdot \mid 0, 1, 0)$, the special case of standard normal distribution. The dashed line is the density of $\mathrm{SN}(\cdot \mid 0, 1, 7)$. The dotted line is the density of $\mathrm{SN}(\cdot \mid 0, 1, \infty)$ corresponding to the special case of a half-normal density.

$f_{U|X}(U \mid X) = \sum_{k=1}^{\infty} \pi_k(X)\{p_k \text{ Normal}(U \mid \mu_{k1}, \sigma_{k1}^2) + (1-p_k) \text{ Normal}(U \mid \mu_{k2}, \sigma_{k2}^2)\}$, with the moment constraint $p_k\mu_{k1} + (1-p_k)\mu_{k2} = 0$ for all k. Use of a two-component mixture of normals as components with each component constrained to have mean zero restricts the mean of the mixture to be zero while allowing the mixture to model other unconstrained aspects of the error distribution. Incorporating covariate information X in modeling the mixture probabilities, this model allows all aspects of the error distribution, other than the mean, to vary nonparametrically with the covariates, not just the conditional variance. Designed for regression problems, these nonparametric models, however, assume that this covariate information is precise. If X is measured with error, as is the case with deconvolution problems, the subject specific residuals may not be informative enough, particularly when the number of replicates per subject is small and the measurement errors have high conditional variability, making simultaneous learning of X and other parameters of the model difficult.

We take a novel semiparametric middle path. The multiplicative structural assumption (2.2) on the measurement errors that reduces the problem of modeling $f_{U|X}$ to two separate problems of modeling a variance function and modeling an error distribution independent of the variable of interest is retained. The difficult problem of flexible modeling of an error distribution with zero mean and unit variance moment restrictions is avoided through a simple reformulation of model (2.2) that replaces the unit variance identifiability restriction on the scaled errors by a similar constraint on the variance function. Model (2.2) is rewritten as

$$U_{ij} = v^{1/2}(X_i)\ \epsilon_{ij} = \frac{v^{1/2}(X_i)}{v^{1/2}(X_0)}\ v^{1/2}(X_0)\epsilon_{ij} = \widetilde{v}^{1/2}(X_i)\ \widetilde{\epsilon}_{ij}, \tag{2.11}$$

where X_0 is arbitrary but fixed point, $\widetilde{v}(X_i) = v(X_i)/v(X_0)$, and $\widetilde{\epsilon}_{ij} = v^{1/2}(X_0)\epsilon_{ij}$. With this specification, $\widetilde{v}(X_0) = 1$, $\mathrm{var}(\widetilde{\epsilon}_{ij}) = v(X_0)$ and $\mathrm{var}(U \mid X) = v(X_0)\widetilde{v}(X)$. The problem of modeling the unrestricted variance function v has now been replaced by the problem of modeling $\widetilde{v}$ restricted to have value 1 at X_0. The problem of modeling the density of ϵ with zero mean and unit variance moment constraints has also been replaced by the easier problem of modeling the density of $\widetilde{\epsilon}_{ij}$ with only a single moment constraint of zero mean.

The conditional variance of the measurement errors is now a scalar multiple of $\widetilde{v}$. So $\widetilde{v}$ can still be referred to as the 'variance function'. The variance of $\widetilde{\epsilon}_{ij}$, however, does not equal unity, but is, in fact, unrestricted. With some abuse of nomenclature, $\widetilde{\epsilon}_{ij}$ is still referred to as the 'scaled errors'. For notational convenience $\widetilde{\epsilon}_{ij}$ is denoted simply by ϵ_{ij}.

The problem of flexibly modeling $\widetilde{v}$ is now addressed. For any X, (i) $b_{q,j}(X) \geq 0\ \forall j$, (ii) $\sum_{j=1}^{J} b_{q,j}(X) = 1$, (iii) $b_{q,j}$ is positive only inside the interval $[t_j, t_{j+q+1}]$, (iv) for $j \in \{(q+1), (q+2), \ldots, (q+K)\}$, for any $X \in (t_j, t_{j+1})$, only $(q+1)$ B-splines $b_{q,j-q}(X), b_{q,j-q+1}(X), \ldots, b_{q,j}(X)$ are positive, and (v) when $X = t_j$, $b_{q,j}(X) = 0$. We let $\widetilde{v}(X) = \mathbf{B}_{q,J}(X)\exp(\boldsymbol{\xi})$, as before, and we use the above mentioned local support properties of the B-spline bases to propose a flexible model for $\widetilde{v}$ subject to $\widetilde{v}(X_0) = 1$. When $X_0 \in (t_j, t_{j+1})$, properties (ii) and (iv) cause the constraint to be simply $\widetilde{v}(X_0) = \sum_{\ell=(q-j)}^{j} b_{q,\ell}(X_0)\exp(\xi_j) = 1$. This is a restriction on only $(q+1)$ of the ξ_j's, and the coefficients of the remaining B-splines remain unrestricted which makes the model for $\widetilde{v}$ very flexible. In a Bayesian framework, the restriction $\widetilde{v}(X_0) = 1$ can be imposed by restricting the support of the prior on $\boldsymbol{\xi}$ to the set

$\{\boldsymbol{\xi} : \sum_{\ell=(q-j)}^{j} b_{q,\ell}(X_0)\exp(\xi_j) = 1\}$. Choosing $X_0 = t_{j_0}$ for some $j_0 \in \{(q+1), \ldots, (q+K)\}$, we further have $b_{j_0}(t_{j_0}) = 0$, and the complete model for $\widetilde{v}$ is given by

$$\begin{aligned}
\widetilde{v}(X) &= \mathbf{B}_{q,J}(X)\exp(\boldsymbol{\xi}), && (2.12)\\
p_0(\boldsymbol{\xi} \mid J, \sigma_\xi^2) &\propto \exp\{-\boldsymbol{\xi}^{\mathrm{T}} P \boldsymbol{\xi}/(2\sigma_\xi^2)\} \times \mathrm{I}\left\{\textstyle\sum_{j=(j_0-q)}^{(j_0-1)} b_{q,j}(t_{j_0})\exp(\xi_j) = 1\right\}, && (2.13)\\
p_0(\sigma_\xi^2) &= \mathrm{IG}(a_\xi, b_\xi), \quad K \sim p_0(K), && (2.14)
\end{aligned}$$

where $\mathrm{I}(\cdot)$ denotes the indicator function.

Now that the variance of ϵ_{ij} has become unrestricted and only a single moment constraint of zero mean is required, a DPMM with mixture components as specified in Pelenis (2014) can be used to model f_ϵ. That is, we let $f_\epsilon(\epsilon) = \sum_{k=1}^{\infty} \pi_{\epsilon k} f_{c\epsilon}(\epsilon \mid p_k, \mu_{k1}, \mu_{k2}, \sigma_{k1}^2, \sigma_{k2}^2)$, $\pi_\epsilon \sim \mathrm{Stick}(\alpha_\epsilon)$, where $f_{c\epsilon}(\epsilon \mid p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \{p\ \mathrm{Normal}(\epsilon \mid \mu_1, \sigma_1^2) + (1-p)\ \mathrm{Normal}(\epsilon \mid \mu_2, \sigma_2^2)\}$, subject to the moment constraint $p\mu_1 + (1-p)\mu_2 = 0$. The moment constraint of zero mean implies that each component density can be described by four parameters. One such parametrization that facilitates prior specification is in terms of parameters $(p, \widetilde{\mu}, \sigma_1^2, \sigma_2^2)$, where (μ_1, μ_2) can be retrieved from $\widetilde{\mu}$ as $\mu_1 = c_1\widetilde{\mu}, \mu_2 = c_2\widetilde{\mu}$, where $c_1 = (1-p)/\{p^2 + (1-p)^2\}^{1/2}$ and $c_2 = -p/\{p^2 + (1-p)^2\}^{1/2}$. Clearly the zero mean constraint is satisfied, since $p\mu_1 + (1-p)\mu_2 = \{pc_1 + (1-p)c_2\}\widetilde{\mu} = 0$. The family includes normal densities as special cases with $(p, \widetilde{\mu}) = (0.5, 0)$ or $(0, 0)$ or $(1, 0)$. Symmetric component densities are obtained as special cases when $p = 0.5$ or $\widetilde{\mu} = 0$. The mixture is symmetric when the all components are as well. Specification of the prior for f_ϵ is completed assuming non-informative priors for $(p, \widetilde{\mu}, \sigma_1^2, \sigma_2^2)$. Letting $\mathrm{Unif}(\ell, u)$ denote a uniform distribution on the interval (ℓ, u), the complete DPMM prior on f_ϵ can then be specified as

$$\begin{aligned}
f_\epsilon(\epsilon) &= \textstyle\sum_{k=1}^{\infty} \pi_{\epsilon k}\ f_{c\epsilon}(\epsilon \mid p_k, \widetilde{\mu}_k, \sigma_{k1}^2, \sigma_{k2}^2), \qquad \pi_\epsilon \sim \mathrm{Stick}(\alpha_\epsilon), && (2.15)\\
(p_k, \widetilde{\mu}_k, \sigma_{k1}^2, \sigma_{k2}^2) &\sim \mathrm{Unif}(0,1)\ \mathrm{Normal}(0, \sigma_{\widetilde{\mu}}^2)\ \mathrm{IG}(a_\epsilon, b_\epsilon)\ \mathrm{IG}(a_\epsilon, b_\epsilon). && (2.16)
\end{aligned}$$

2.3 Model Diagnostics

In practical deconvolution problems, the basic structural assumptions on the measurement errors may be dictated by prominent features of the data extracted by simple diagnostic tools and expert knowledge of the data generating process. Conditional heteroscedasticity, in particular, is easy to identify from the scatterplot of S_W^2 on $\overline{W}$, where $\overline{W}$ and S_W^2 denote the subject specific sample mean and variance, (Eck-

ert, et al., 1997). The multiplicative structural assumption (2.2) on the measurement errors provides one particular way of accommodating conditional heteroscedasticity in the model. When at least 4 replicates are available for each subject, one can define the pairs $(W_{ij_1}, C_{ij_2j_3j_4})$ for all i and for all $j_1 \neq j_2 \neq j_3 \neq j_4$, where $C_{ij_2j_3j_4} = \{(W_{ij_2} - W_{ij_3})/(W_{ij_2} - W_{ij_4})\}$. When (2.2) is true, $C_{j_2j_3j_4} = \{(\epsilon_{j_2} - \epsilon_{j_3})/(\epsilon_{j_2} - \epsilon_{j_4})\}$ is independent of W_{j_1}. Therefore, absence of nonrandom patterns in the plots of W_{j_1} against $C_{j_2j_3j_4}$ and nonsignificant p-values in nonparametric tests of association between W_{j_1} and $C_{j_2j_3j_4}$ for various $j_1 \neq j_2 \neq j_3 \neq j_4$ may be taken as indications that (2.2) is valid or that the departures from (2.2) are not severe. If there are m (≥ 4) replicates per subject, the total number of possible such tests is $m!/(m-4)! = L$, say, where, for any positive integer r, $r! = r \cdot (r-1) \ldots 2 \cdot 1$. The p-values of these tests can be combined using the truncated product method of Zaykin, et al. (2002). The test statistic of this combined left-sided test is given by $T(\varsigma) = \prod_{\ell=1}^{L} p_\ell^{1(p_\ell < \varsigma)}$, where p_ℓ denotes the p-value of the ℓ^{th} test and ς is a prespecified truncation limit. If $\min_\ell\{p_\ell\} \geq \varsigma$, the p-value of the combined test is trivially 1. Otherwise, the bootstrap procedure described in Zaykin, et al. (2002) may be used to estimate it.

2.4 Choice of Hyper-Parameters

For the DPMM prior for f_X, the prior variance of each σ_k^2 is $\sigma_0^4/\{(\gamma_0-2)^2(\gamma_0-1)\}$, whereas the prior variance of each μ_k, given σ_k^2, is σ_k^2/ν_0. Small values of γ_0 and ν_0 imply large prior variance and hence non-informativeness. We chose $\gamma_0 = 3$ and $\nu_0 = 1/5$. The prior marginal mean and variance of X, obtained by integrating out all but the hyper-parameters, are given by μ_0 and $\sigma_0^2(1+1/\nu_0)/(\gamma_0-1)$ respectively. Taking an empirical Bayes type approach, we set $\mu_0 = \overline{\mathbf{W}}$ and $\sigma_0^2 = S_{\mathbf{W}}^2(\gamma_0-1)/(1+1/\nu_0)$, where $\overline{\mathbf{W}}$ is the mean of the subject-specific sample means $\overline{\mathbf{W}}_{1:n}$, and $S_{\mathbf{W}}^2$ is an estimate of the across subject variance from a one way random effects model. To ensure noninformativeness, hyper-parameters appearing in the prior for f_ϵ are chosen as $\sigma_{\widetilde{\mu}} = 3$, $a_\epsilon = 1$ and $b_\epsilon = 1$. For real world applications, the values of A and B may not be known. We set $[A, B] = [\min(\overline{\mathbf{W}}_{1:n}) - 0.1 \text{ range}(\overline{\mathbf{W}}_{1:n}), \max(\overline{\mathbf{W}}_{1:n}) + 0.1 \text{ range}(\overline{\mathbf{W}}_{1:n})]$. The DP concentration parameters α_X and α_ϵ could have been assigned gamma hyper-priors (Escobar and West, 1995), but in this dissertation we kept them fixed at $\alpha_X = 0.1$ and $\alpha_\epsilon = 1$, respectively. The prior mean and standard deviation of λ were set at $\mu_{0\lambda} = 0$ and $\sigma_{0\lambda} = 4$. For modeling the variance functions v and $\widetilde{v}$, quadratic (q=2) B-splines based are used. See Appendix A.3 for detailed

expressions. The B-splines are based on $(2 \times 2 + 10 + 1) = 15$ knot points that divide the interval $[A, B]$ into $K = 10$ subintervals of equal length. We take $X_0 = t_5$. The identifiability restriction on the variance function for Model III now becomes $\{\exp(\xi_3)+\exp(\xi_4)\} = 2$. The inverse-gamma hyper-prior on the smoothing parameter σ_ξ^2 is non-informative if b_ξ is small relative to $\boldsymbol{\xi}^{\mathrm{T}} P \boldsymbol{\xi}$. We chose $a_\xi = b_\xi = 0.1$.

2.5 Posterior Inference

Define cluster labels $\mathbf{C}_{1:n}$, where $C_i = k$ if X_i is associated with the k^{th} component of the DPMM. Similarly for Model-III, define cluster labels $\{Z_{ij}\}_{i,j=1}^{n,m_i}$, where $Z_{ij} = k$ if ϵ_{ij} comes from the k^{th} component of (2.15). Let $N = \sum_{i=1}^{n} m_i$ denote the total number of observations. With a slight abuse of notation, define $\mathbf{W}_{1:N} = \{W_{ij}\}_{i,j=1}^{n,m_i}$ and $\mathbf{Z}_{1:N} = \{Z_{ij}\}_{i,j=1}^{n,m_i}$. Then for Model-I, $f_{W|X}(W_{ij} \mid X_i, \boldsymbol{\xi}) = \text{Normal}\{W_{ij} \mid X_i, v(X_i, \boldsymbol{\xi})\}$; for Model-II, $f_{W|X}(W_{ij} \mid X_i, \boldsymbol{\xi}, \lambda) = \text{SN}\{W_{ij} \mid X_i, v(X_i, \boldsymbol{\xi}), \lambda\}$; and for Model-III, given $Z_{ij} = k$, $f_{W|X}(W_{ij} \mid X_i, \boldsymbol{\xi}, p_k, \mu_{k1}, \mu_{k2}, \sigma_{k1}^2, \sigma_{k2}^2) = p_k \text{Normal}\{W_{ij} \mid X_i + \widetilde{v}(X_i, \boldsymbol{\xi})^{1/2}\mu_{k1}, \widetilde{v}(X_i, \boldsymbol{\xi})\sigma_{k1}^2\} + (1-p_k)\text{Normal}\{W_{ij} \mid X_i + \widetilde{v}(X_i, \boldsymbol{\xi})^{1/2}\mu_{k2}, \widetilde{v}(X_i, \boldsymbol{\xi})\sigma_{k2}^2\}$. In what follows $\boldsymbol{\zeta}$ denotes a generic variable that collects all other parameters of a model, including $\mathbf{X}_{1:n}$, that are not explicitly mentioned.

Inference is based on samples drawn from the posterior using Markov chain Monte Carlo techniques. It is possible to integrate out the random mixture probabilities from the prior and posterior full conditionals of the cluster labels. Classical algorithms for fitting DPMMs make use of this and work with the resulting Polya urn scheme. Neal (2000) provided an excellent review of this type of algorithm for both conjugate and non-conjugate cases. We update the parameters specific to DPMMs using algorithms specific to those models and other parameters are updated using the Metropolis-Hastings algorithm. In what follows, the generic notation $q(\textit{current} \to \textit{proposed})$ denotes the proposal distributions of the Metropolis-Hastings steps proposing a move from the *current* value to the *proposed* value.

The starting values of the MCMC chain are determined as follows. Subject-specific sample means $\overline{\mathbf{W}}_{1:n}$ are used as starting values for $\mathbf{X}_{1:n}$. Each C_i is initialized at i with each X_i coming from its own cluster with mean $\mu_i = X_i$ and variance $\sigma_i^2 = \sigma_0^2$. In addition, σ_ξ^2 is initialized at 0.1. The initial value of $\boldsymbol{\xi}$ is obtained by maximizing $\ell(\boldsymbol{\xi} \mid 0.1, \overline{\mathbf{W}}_{1:n})$ with respect to $\boldsymbol{\xi}$, where $\ell(\boldsymbol{\xi} \mid \sigma_\xi^2, \mathbf{X}_{1:n})$ denotes the conditional log-posterior of $\boldsymbol{\xi}$. The parameters of the distribution of scaled errors are initialized at values that correspond to the special standard normal case. For

example, for Model-II, λ is initialized at zero. For Model-III, Z_{ij}'s are all initialized at 1 with $(p_1, \widetilde{\mu}_1, \sigma_{11}^2, \sigma_{12}^2) = (0.5, 0, 1, 1)$. The MCMC iterations comprise the following steps.

1. **Updating the parameters of the distribution of X:** Conditionally given $\mathbf{X}_{1:n}$, the parameters of the DPMM for f_X can be updated using a Gibbs sampler (Neal, 2000, Algorithm 2). The full conditional of C_i is given by

$$\begin{aligned} p(C_i = k, k \in \mathbf{C}_{-i} \mid \mathbf{X}_{1:n}, \mathbf{C}_{-i}, \boldsymbol{\zeta}) &= b \frac{n_{-i,k}}{n - 1 + \alpha_X} \text{ Normal}(X_i \mid \mu_k, \sigma_k^2), \\ p(C_i \notin \mathbf{C}_{-i} \mid \mathbf{X}_{1:n}, \mathbf{C}_{-i}, \boldsymbol{\zeta}) &= b \frac{\alpha_X}{n - 1 + \alpha_X} \text{ t}_{2\gamma_0}(t_i), \end{aligned}$$

   where b denotes the appropriate normalizing constant; for each i, $\mathbf{C}_{-i} = \mathbf{C}_{1:n} - \{C_i\}$; $n_{-i,k} = \sum_{\{l:l \neq i\}} 1_{\{c_l = k\}}$ is the number of c_l's that equal k in $\mathbf{C}_{-i}$; and $t_i = \gamma_0^{1/2}(X_i - \mu_0)/\{\sigma_0(1 + 1/\nu_0)^{1/2}\}$. t_m denotes the density of a t-distribution with m degrees of freedom.

   For all $k \in \mathbf{C}_{1:n}$, we update (μ_k, σ_k^2) using the closed-form joint full conditional given by $\{(\mu_k, \sigma_k^2) \mid \mathbf{X}_{1:n}, \boldsymbol{\zeta}\} = \text{NIG}(\mu_{nk}, \sigma_{nk}^2/\nu_{nk}, \gamma_{nk}, \sigma_{nk}^2)$, where $n_k = \sum_{i=1}^n 1_{\{C_i = k\}}$ is the number of X_i's associated with the k^{th} cluster; $\nu_{nk} = (\nu_0 + n_k)$; $\gamma_{nk} = (\gamma_0 + n_k/2)$; $\mu_{nk} = (\nu_0 \mu_0 + n_k \sum_{\{i:C_i = k\}} X_i)/(\nu_0 + n_k)$ and $\sigma_{nk}^2 = \sigma_0^2 + (\sum_{\{i:C_i = k\}} X_i^2 + \nu_0 \mu_0^2 - \nu_{nk} \mu_{nk}^2)/2$.

2. **Updating $\mathbf{X}_{1:n}$:** Because the X_i's are conditionally independent, the full conditional of X_i is given by $p(X_i \mid \mathbf{W}_{1:N}, \boldsymbol{\zeta}) \propto \widehat{f}_X(X_i \mid \boldsymbol{\zeta}) \times \prod_{j=1}^{m_i} f_{W|X}(W_{ij} \mid X_i, \boldsymbol{\zeta})$. We use a Metropolis-Hastings sampler to update the X_i's with proposal $q(X_i \to X_{i,new}) = TN(X_{i,new} \mid X_i, \sigma_X^2, [A, B])$, where $\sigma_X =$ (range of $\overline{\mathbf{W}}_{1:n}$)/6 and $\text{TN}(\cdot \mid m, s^2, [\ell, u])$ denotes a truncated normal distribution with location m and scale s restricted to the interval $[\ell, u]$.

3. **Updating the parameters of the distribution of scaled errors:** For Model-II and Model-III, the parameters involved in the distribution of scaled errors have to be updated.

   For Model-II, the distribution of scaled error is $\text{SN}(0, 1, \lambda)$, involving only the parameter λ. The full conditional of λ is given by $p(\lambda \mid \mathbf{W}_{1:N}, \boldsymbol{\zeta}) \propto p_0(\lambda) \times \prod_{i=1}^n \prod_{j=1}^{m_i} f_{W|X}(W_{ij} \mid \lambda, \boldsymbol{\zeta})$. We use Metropolis-Hastings sampler to update λ with random walk proposal $q(\lambda \to \lambda_{new}) = \text{Normal}(\lambda_{new} \mid \lambda, \sigma_\lambda^2)$.

For Model-III, we use Metropolis-Hastings samplers to update the latent parameters $\mathbf{Z}_{1:N}$ as well as the component specific parameters $(p_k, \widetilde{\mu}_k, \sigma^2_{k1}, \sigma^2_{k2})$'s (Neal, 2000, Algorithm 5). We propose a new value of Z_{ij}, say $Z_{ij,new}$, according to its marginalized conditional prior

$$\begin{aligned} p(Z_{ij} = k, k \in \mathbf{Z}_{-ij} \mid \mathbf{Z}_{-ij}) &= N_{-ij,k}/(N - 1 + \alpha_\epsilon), \\ p(Z_{ij} \notin \mathbf{Z}_{-ij} \mid \mathbf{Z}_{-ij}) &= \alpha_\epsilon/(N - 1 + \alpha_\epsilon), \end{aligned}$$

where, for each (i, j) pair, $\mathbf{Z}_{-ij} = \mathbf{Z}_{1:N} - \{Z_{ij}\}$; $N_{-ij,k} = \sum_{\{rs:rs \neq ij\}} 1_{\{Z_{rs}=k\}}$, the number of Z_{rs}'s in $\mathbf{Z}_{-ij}$ that equal k. If $Z_{ij,new} \notin \mathbf{Z}_{-ij}$, we draw a proposed value $(p_{Z_{ij,new}}, \widetilde{\mu}_{Z_{ij,new}}, \sigma^2_{Z_{ij,new}1}, \sigma^2_{Z_{ij,new}2})$ from the prior $p_0(p, \widetilde{\mu}, \sigma^2_1, \sigma^2_2)$. We update Z_{ij} to its proposed value with probability

$$\min\left\{1, \frac{f_{W|X}(W_{ij} \mid p_{Z_{ij,new}}, \widetilde{\mu}_{Z_{ij,new}}, \sigma^2_{Z_{ij,new}1}, \sigma^2_{Z_{ij,new}2}, \boldsymbol{\zeta})}{f_{W|X}(W_{ij} \mid p_{Z_{ij}}, \widetilde{\mu}_{Z_{ij}}, \sigma^2_{Z_{ij}1}, \sigma^2_{Z_{ij}2}, \boldsymbol{\zeta})}\right\}.$$

For all $k \in \mathbf{Z}_{1:N}$, we propose a new value for $(p_k, \widetilde{\mu}_k, \sigma^2_{k1}, \sigma^2_{k2})$ with the proposal

$$\begin{aligned} q\{\boldsymbol{\theta}_k = (p_k, \widetilde{\mu}_k, \sigma^2_{k1}, \sigma^2_{k2}) \to (p_{k,new}, \widetilde{\mu}_{k,new}, \sigma^2_{k1,new}, \sigma^2_{k2,new}) = \boldsymbol{\theta}_{k,new}\} = \\ \mathrm{TN}(p_{k,new} \mid p_k, \sigma^2_p, [0,1]) \ \times \ \mathrm{Normal}(\widetilde{\mu}_{k,new} \mid \widetilde{\mu}_k, \sigma^2_{\widetilde{\mu}}) \ \times \\ \mathrm{TN}(\sigma^2_{k1,new} \mid \sigma^2_{k1}, \sigma^2_\sigma, [\max\{0, \sigma^2_{k1} - 1\}, \sigma^2_{k1} + 1]) \ \times \\ \mathrm{TN}(\sigma^2_{k2,new} \mid \sigma^2_{k2}, \sigma^2_\sigma, [\max\{0, \sigma^2_{k2} - 1\}, \sigma^2_{k2} + 1]). \end{aligned}$$

We update $\boldsymbol{\theta}_k$ to the proposed value $\boldsymbol{\theta}_{k,new}$ with probability

$$\min\left\{1, \frac{q(\boldsymbol{\theta}_{k,new} \to \boldsymbol{\theta}_k)}{q(\boldsymbol{\theta}_k \to \boldsymbol{\theta}_{k,new})} \frac{\prod_{\{ij:z_{ij}=k\}} f_{W|X}(W_{ij} \mid \boldsymbol{\theta}_{k,new}, \boldsymbol{\zeta}) \ p_0(\boldsymbol{\theta}_{k,new})}{\prod_{\{ij:z_{ij}=k\}} f_{W|X}(W_{ij} \mid \boldsymbol{\theta}_k, \boldsymbol{\zeta}) \ p_0(\boldsymbol{\theta}_k)}\right\}.$$

4. **Updating the parameters of the variance function:** The full conditional for $\boldsymbol{\xi}$ is given by $p(\boldsymbol{\xi} \mid \mathbf{W}_{1:N}, \boldsymbol{\zeta}) \propto p_0(\boldsymbol{\xi}) \times \prod_{i=1}^n \prod_{j=1}^{m_i} f_{W|X}(W_{ij} \mid \boldsymbol{\xi}, \boldsymbol{\zeta})$. We use Metropolis-Hastings sampler to update $\boldsymbol{\xi}$ with random walk proposal $q(\boldsymbol{\xi} \to \boldsymbol{\xi}_{new}) = \mathrm{MVN}_J(\boldsymbol{\xi}_{new} \mid \boldsymbol{\xi}, \Sigma_\xi)$. Here $\mathrm{MVN}_J(\boldsymbol{\mu}, \Sigma)$ denotes a J-variate normal distribution with mean $\boldsymbol{\mu}$ and positive semidefinite covariance matrix Σ. For Model III, the identifiability restriction is imposed by replacing $\xi_{new,3} = \log\{2 - \exp(\xi_{new,4})\}$.

Finally, we update the hyper-parameter σ_ξ^2 using its closed-form full conditional $(\sigma_\xi^2 \mid \boldsymbol{\xi}, \boldsymbol{\zeta}) = \text{IG}\{a_\xi + (J+2)/2, b_\xi + \boldsymbol{\xi}^{'} P \boldsymbol{\xi}/2\}$.

The covariance matrix Σ_ξ of the proposal distribution for $\boldsymbol{\xi}$ is taken to be the inverse of the negative Hessian matrix of $l(\boldsymbol{\xi} \mid 0.1, \overline{\mathbf{W}}_{1:n})$ evaluated at the chosen initial value of $\boldsymbol{\xi}$. See Appendix A.2 for more details. Other variance parameters appearing in the proposal distributions are tuned to get good acceptance rates for the Metropolis-Hastings samplers, the values $\sigma_\lambda = 1, \sigma_p = 0.01$ and $\sigma_\sigma = 0.1$ working well in the examples considered. In simulation experiments, 5,000 MCMC iterations with the initial 3,000 discarded as burn-in produced very stable estimates of the density and the variance function.

The posterior estimate of f_X is given by the unconditional predictive density $f_X(\cdot \mid \mathbf{W}_{1:N})$. A Monte Carlo estimate of $f_X(\cdot \mid \mathbf{W}_{1:N})$, based on M samples from the posterior, is given by

$$\widehat{f}_X(X \mid \mathbf{W}_{1:N}) = M^{-1} \sum_{m=1}^{M} \Bigg[\sum_{k=1}^{k^{(m)}} \{n_k^{(m)}/(\alpha_X + n)\} \text{ Normal}(X \mid \mu_k^{(m)}, \sigma_k^{(m)2}) + \{\alpha_X/(\alpha_X + n)\} \text{ t}_{2\gamma_0}(t_X) \Bigg],$$

where $t_X = t(X) = \gamma_0^{1/2}(X - \mu_0)/\{\sigma_0(1 + 1/\nu_0)^{1/2}\}$, $(\mu_k^{(m)}, \sigma_k^{(m)2})$ is the sampled value of (μ_k, σ_k^2) in the m^{th} sample, $n_k^{(m)}$ is the number of X_i's associated with the k^{th} cluster, and $k^{(m)}$ is the total number of active clusters. With $(p_k^{(m)}, \widetilde{\mu}_k^{(m)}, \sigma_{k1}^{(m)2}, \sigma_{k2}^{(m)2})$, $N_k^{(m)}$ and $k_\epsilon^{(m)}$ defined in a similar fashion, the posterior Monte Carlo estimate of f_ϵ for Model-III is

$$\widehat{f}_\epsilon(\epsilon \mid \mathbf{W}_{1:N}) = M^{-1} \sum_{m=1}^{M} \Bigg[\sum_{k=1}^{k_\epsilon^{(m)}} \{N_k^{(m)}/(\alpha_\epsilon + N)\} \; f_{c\epsilon}(\epsilon \mid p_k^{(m)}, \widetilde{\mu}_k^{(m)}, \sigma_{k1}^{(m)2}, \sigma_{k2}^{(m)2}) + \{\alpha_\epsilon/(\alpha_\epsilon + N)\} \int f_{c\epsilon}(\epsilon \mid p, \widetilde{\mu}, \sigma_{k1}^2, \sigma_{k2}^2) dp_0(p, \widetilde{\mu}, , \sigma_{k1}^2, \sigma_{k2}^2) \Bigg],$$

The integral above can not be exactly evaluated. Monte Carlo approximation may be used. If $N >> \alpha_\epsilon$, the term may simply be neglected. For Model II, f_ϵ can be estimated by $\widehat{f}_\epsilon(\epsilon \mid \mathbf{W}_{1:N}) = \sum_{m=1}^{M} \text{SN}(\epsilon \mid 0, 1, \lambda^{(m)})/M$. For Models I and II, an estimate of the variance function v can similarly be obtained as $\widehat{v}(X \mid \mathbf{W}_{1:N}) = \sum_{m=1}^{M} v(X \mid \boldsymbol{\xi}^{(m)})/M$. An estimate of the restricted variance function $\widetilde{v}$ for Model III can be obtained using a similar formula. For Model III, $\widehat{v}$ and a scaled version of

$\widehat{f_\epsilon}$, scaled to have unit variance, can be obtained using the estimate of $\widetilde{v}(X_0)$.

The mean integrated squared error (MISE) of estimation of f_X by $\widehat{f_X}$ is defined as $MISE = E\int\{f_X(X) - \widehat{f_X}(X)\}^2 dX$. Based on B simulated data sets, a Monte Carlo estimate of MISE is given by $MISE_{est} = B^{-1}\sum_{b=1}^{B}\sum_{i=1}^{N}\{f_X(X_i^\Delta) - \widehat{f}_X^{(b)}(X_i^\Delta)\}^2\Delta_i$, where $\{X_i^\Delta\}_{i=0}^N$ are a set of grid points on the range of X and $\Delta_i = (X_i^\Delta - X_{i-1}^\Delta)$ for all i.

2.6 Simulation Experiments

Simulation experiments are designed to evaluate the MISE performance of the proposed models for a wide range of possibilities. The deconvolution models proposed in this dissertation all take semiparametric routes to model conditional heteroscedasticity assuming a multiplicative structural assumption on the measurement errors. Performance of the proposed models is first evaluated for 'semiparametric truth scenarios' when the truth conforms to the assumed multiplicative structure. Efficiency of the proposed models will also be illustrated for 'nonparametric truth' scenarios when the truth departs from the assumed multiplicative structure.

The reported estimated MISEs are all based on $B = 400$ simulated data sets. For the proposed methods 5,000 MCMC iterations were run in each case with the initial 3,000 iterations discarded as burn-in. We programmed in R. With $n = 500$ subjects and $m_i = 3$ proxies for each subject, on an ordinary desktop, $5,000$ MCMC iterations for models I, II and III required approximately 5 minutes, 10 minutes and 25 minutes, respectively. In comparison, the method of Staudenmayer, et al. (2008) and the nonparametric alternative described in Section 2.6.2 took approximately 100 minutes and 150 minutes, respectively.

2.6.1 Semiparametric Truth

This subsection presents the results of simulation experiments comparing our methods with the method of Staudenmayer, et al. (2008). Their approach, referred to as the SRB method henceforth, has been shown previously to out-perform the deconvoluting kernel-based method of Delaigle and Meister (2008). The methods are compared over a factorial combination of three sample sizes ($n = 250, 500, 1000$), two densities for X $\{f_X^1(X) = 0.5\ \text{Normal}(X \mid 0, 0.75) + 0.5\ \text{Normal}(X \mid 3, 0.75)$ and $f_X^2(X) = 0.8\ \text{Normal}(X \mid 0, 0.75) + 0.2\ \text{Normal}(X \mid 3, 0.75)\}$, nine different types of distributions for the scaled errors (six light-tailed and three heavy-tailed, see Table 2.1 and Figure 2.3), and one variance function $v(X) = (1 + X/4)^2$. For each subject,

$m_i = 3$ replicates were simulated.

2.6.1.1 Results for Light-tailed Error Distributions

This section discusses MISE performances of the models for the 36 ($3 \times 2 \times 6$) cases where the distribution of the scaled errors were light-tailed, distributions (a)-(f), see Table 2.1 and Figure 2.3. The MISEs are presented in Table 2.2. Results of the simulation experiments show that all three models proposed in this dissertation significantly out-performed the SRB model in all 36 cases considered. When measurement errors are normally distributed, the reductions in MISE over the SRB method for all three models and for all six possible combination of sample sizes and true X distributions are more than 50%. This is particularly interesting, since the SRB method was originally proposed for normally distributed errors, even more so because our Model-II and Model-III relax the normality assumption on the measurement errors.

2.6.1.2 Results for Heavy-tailed Error Distributions

This section discusses MISE performances of the models for the 18 ($3 \times 2 \times 3$) cases where the distribution of scaled errors were heavy-tailed, distributions (g), (h) and (i), see Table 2.1 and Figure 2.3. The MISEs are presented in Table 2.3. Results for the error distributions (g) and (h) are summarized in Figure 2.4 and Figure 2.5, respectively. The SRB model and Model-I assume normally distributed errors; Model-II assumes skew-normal errors whose tail behavior is similar to that of normal distributions. The results show the MISE performances of these three models to be very poor for heavy-tailed error distributions and the MISEs increased with an increase in sample size due to the presence of an increasing number of outliers. Model-III, on the other hand, could accommodate heavy-tails in the error distributions and was, therefore, very robust to the presence of outliers. MISE patterns produced by Model-III for heavy-tailed errors were similar to that for light-tailed errors, and improvements in MISEs over the other models were huge. For example, when the density for the scaled was (i), a mixture of Laplace densities with a very sharp peak at zero, for $n = 1000$, the improvements in MISEs over the SRB model were $54.03/0.94 \approx 57$ times for the 50-50 mixture of normals and $57.87/0.83 \approx 70$ times for the 80-20 mixture of normals.

2.6.2 Nonparametric Truth

This subsection is aimed at providing some empirical support to the claim made in Section 2.2.3.3, where it was argued that for deconvolution problems the proposed semiparametric route to model the distribution of conditionally heteroscedastic measurement errors would often be more efficient than possible nonparametric alternatives, even when the truth departs from the assumed multiplicative structural assumption (2.2) on the measurement errors. This is done by comparing our Model III with a method that also models the density of interest by a DPMM like ours but employs the formulation of Pelenis (2014) to model the density of the measurement errors. This possible nonparametric alternative was reviewed in Section 2.2.3.3 and will be referred to as the NPM method henceforth. Recall that by modeling the mixture probabilities as functions of X the NPM model allows all aspects of the distribution of errors to vary with X, not just the conditional variance. In theory, the NPM model is, therefore, more flexible than Model-III as it can also accommodate departures from (2.2). However, in practice, for reasons described in Section 2.2.3.3, Model-III will often be more efficient than the NPM model, as is shown here.

In the simulation experiments the true conditional distributions that generate the measurement errors are designed to be of the form $f_{U|X}(U \mid X) = \sum_{k=1}^{K} \pi_k(X) f_{cU}(U \mid \sigma^2_{Uk}, \boldsymbol{\theta}_{Uk})$, where each component density has mean zero, the k^{th} component has variance σ^2_{Uk}, and $\boldsymbol{\theta}_{Uk}$ denotes additional parameters. For the true and the fitted mixture probabilities we used the formulation of Chung and Dunson (2009) that allows easy posterior computation through data augmentation techniques. That is, we took $\pi_k(X) = V_k(X) \prod_{\ell=1}^{k-1}\{1 - V_\ell(X)\}$ with $V_k(X) = \Phi(\alpha_k - \beta_k |X - X^\star_k|)$ for $k = 1, 2, \ldots, (K-1)$ and $\pi_K(X) = \{1 - \sum_{k=1}^{K-1} \pi_k(X)\}$. The truth closely resembles the NPM model and clearly departs from the assumptions of Model III. The conditional variance is now given by $\text{var}(U \mid X) = \sum_{k=1}^{K} \pi_k(X)\sigma^2_{UK}$. The two competing models are then compared over a factorial combination of three sample sizes ($n = 250, 500, 1000$), two densities for X - f^1_X and f^2_X, as defined in Section 2.6.1, and three different choices for the component densities f_{cU} - (j) Normal$(0, \sigma^2_{Uk})$, (k) SN$(\cdot \mid 0, \sigma^2_{Uk}, \lambda_U)$ and (l) $SN(\cdot \mid 0, \sigma^2_{Uk}, \lambda_{Uk})$. In each case, $K = 8$ and the parameters specifying the true mixture probabilities are set at $\alpha_k = 2$, $\beta_k = 1/2$ for all k with $X^\star_k$ taking values in $\{-1.9, -1, 0, 1, 2.5, 4, 5.5\}$ in that order. We chose the priors for α_k, β_k and $X^\star_k$ as in Chung and Dunson (2009). The component specific variance parameters σ^2_{Uk} are set by minimizing the sum of squares of $g(X) = \{(1 + X/4)^2 -$

$\sum_{k=1}^{K} \pi_k(X)\sigma_{Uk}^2\}$ on a grid. For the density (k) we set $\lambda_U = 7$. For the density (l) λ_{Uk} take values in $\{7, 3, 1, 0, -1, -3, -7\}$, with λ_{Uk} decreasing as X increases. For each subject, $m_i = 3$ replicates were simulated.

The estimated MISEs are presented in Table 2.4. The results show that Model III vastly outperforms the NPM model in all 18 ($3 \times 2 \times 3$) cases even though the truth actually conforms to the NPM model closely. The reductions in the MISEs are particularly significant when the true density of interest is a 50-50 mixture of normals. The results further emphasize the need for flexible and efficient semiparametric deconvolution models such as the ones proposed in this dissertation.

2.7 Application in Nutritional Epidemiology

2.7.1 Data Description and Model Validation

Dietary habits are known to be leading causes of many chronic diseases. Accurate estimation of the distributions of dietary intakes is important in nutritional epidemiologic surveillance and epidemiology. One large scale epidemiologic study conducted by the National Cancer Institute (NCI), the Eating at America's Table (EATS) study (Suber, et al., 2001), serves as the motivation for this section. In this study $n = 965$ participants were interviewed $m_i = 4$ times over the course of a year and their 24 hour dietary recalls (W_{ij}'s) were recorded. The goal is to estimate the distribution of true daily intakes (X_i's).

Figure 2.6 shows diagnostic plots (as described in Section 2.3) for daily intakes of folate. Conditional heteroscedasticity of measurements errors is one salient feature of the data, clearly identifiable from the plot of subject-specific means versus subject-specific variances. The authors did not see any nonrandom pattern in the scatterplots of W_{j_1} vs $C_{j_2 j_3 j_4}$ for various $j_1 \neq j_2 \neq j_3 \neq j_4$. Combined p-value of 1 given by nonparametric tests of association combined by the truncated product method of Zaykin, et al. (2002) with truncation limit as high as 0.50 is also strong evidence in favor of independence of W_{j_1} and $C_{j_2 j_3 j_4}$ for all $j_1 \neq j_2 \neq j_3 \neq j_4$. By the arguments presented in Section 2.3, model (1)-(2) may therefore be assumed to be valid for reported daily intakes of folate. Data on many more dietary components were recorded in the EATS study. Due to space constraints, it is not possible to present diagnostic plots for other dietary components. However, it should be noted that the combined p-values for nonparametric tests of association between W_{j_1} and $C_{j_2 j_3 j_4}$ for various $j_1 \neq j_2 \neq j_3 \neq j_4$ for *all* 25 dietary components, for which daily

dietary intakes were recorded in the EATS study, are greater than 0.50 even for truncation limit as high as 0.50. See Table 2.5.

2.7.2 Results for Daily Intakes of Folate

Estimates of the density of daily intakes of folate and other nuisance functions of secondary importance produced by different deconvolution models are summarized in Figure 2.7. When the density of scaled errors is allowed to be flexible, as in Model-III, the estimated density of daily folate intakes is visibly very different from the estimates when the measurement errors are assumed to be normally or skew-normally distributed, as in Model-I, Model-II or the SRB model, particularly in the interval of 3-6 mcg. Estimated 90% credible intervals for $f_X(3.7)$ for Model-I is $(0.167, 0.283)$, for Model-II is $(0.237, 0.375)$, and for Model-III is $(0.092, 0.163)$. Since the credible interval for Model-III is disjoint from the credible intervals for the other models, the differences in the estimated densities at 3.7 may be considered to be significant.

Our analysis also showed that the measurement error distributions of *all* dietary components included in the EATS study deviate from normality and exhibit strong conditional heteroscedasticity. These findings emphasize the importance of flexible conditionally heteroscedastic error distribution models in nutritional epidemiologic studies.

2.8 Conclusion

2.8.1 Summary

We have considered the problem of Bayesian density deconvolution in the presence of conditionally heteroscedastic measurement errors. Attending to the specific needs of deconvolution problems, three different approaches were considered for modeling the distribution of measurement errors. The first model made the conventional normality assumption about the measurement errors. The next two models allowed, with varying degrees of flexibility, the distribution of measurement errors to deviate from normality. In all these models conditional heteroscedasticity was also modeled nonparametrically. The proposed methodology, therefore, makes important contributions to the density deconvolution literature, allowing both the distribution of interest and the distribution of measurement errors to deviate from standard parametric laws, while at the same time accommodating conditional heteroscedasticity. Efficiency of the models in recovering the true density of interest was illustrated

through simulation experiments, and in particular we showed that our method vastly dominates that of Staudenmayer, et al. (2008). Results of the simulation experiments suggested that all the models introduced in this dissertation out-perform previously existing methods, even while relaxing some of the restrictive assumptions of previous approaches. Simulation experiments also showed that our Bayesian semiparametric deconvolution approaches proposed in this dissertation would often be more efficient than possible nonparametric alternatives, even when the true data generating process deviates from the assumed semiparametric framework.

2.8.2 Data Transformation and Homoscedasticity

In our application area of nutrition, many researchers assume that W is unbiased for X in the original scale that the nutrient is measured, i.e., $E(W|X) = X$ as in our model, see Willett (2012), Spiegelman, et al. (1997, 2001, 2005) and Kipnis, et al. (2009). It is this original scale of X then that is of scientific interest in this instance. An alternative technique is a transform-retransform method: attempt to transform the W_{ij} data to make it additive and with homoscedastic measurement error, fit in the transformed scale, and then back-transform the density. For example, if $W_{ij} = X_i \exp(U_{ij} - \sigma_u^2/2)$ where $U_{ij} = \text{Normal}(0, \sigma_u^2)$, then $\log(W_{ij}) = \log(X_i) - \sigma_u^2/2 + U_{ij}$, the classical homoscedastic deconvolution problem with target $X_* = \log(X) - \sigma_u^2/2$. One could then use any homoscedastic deconvolution method to estimate the density of X_*, and then from that estimate the density of X. Our methods obviously apply to such a problem. We have used the kernel deconvolution R package "decon" (Wang and Wang, 2011), the only available set of programs, and compared it to our method both using transform-retransform with homoscedasticity and by working in the original scale, using Model III. In a variety of target distributions for X and a variety of sample sizes, our methods consistently have substantially lower MISE.

It is also the case though that transformations to a model such as $h(W) = h(X) + U$ with $U = \text{Normal}(0, \sigma_u^2)$ do not satisfy the unbiasedness condition in the original scale. In the log-transformation case, there is a multiplicative bias, but in the cube-root case, $E(W) = E(X) + 3\sigma_u^2 E(X^{1/3})$, a model that many in nutrition would find uncomfortable and, indeed, objectionable.

Of course, other fields would be amenable to unbiasedness on a transformed scale, and hope that the measurement error is homoscedastic on that scale. Even in this problem, our methodology is novel and dominates other methods that have

been proposed previously. Our methods apply to this problem, allowing flexible Bayesian semiparametric models for the density of X in the transformed scale, flexible Bayesian semiparametric models for the density of the measurement errors, and, if desired, at the same time build modeling robustness lest there be any remaining heteroscedasticity. We have experimented with this ideal case, and even here our methods substantially dominate those currently in the literature. It also must be remembered too that it is often not possible to transform to additivity with homoscedasticity: one example is the EATS data of Section 3.4, where this occurs with vitamin B for the Box-Cox family. Details are provided in Appendix A.

2.8.3 Extensions

Application of the Bayesian semiparametric methodology, introduced in this section for modeling conditionally heteroscedastic errors with unknown distribution where the conditioning variable is not precisely measured, is not limited to deconvolution problems. An important extension of this work and the subject of the next section of the dissertation is an application of the proposed methodology to errors-in-variables regression problems.

Distribution of scaled errors	Skewness (γ_1)	Excess Kurtosis (γ_2)
(a) Normal(0,1)	0	0
(b) Skew-normal(0,1,7)	0.917	0.779
(c) SMRTCN(1,1,0.4,2,2,1)	0.499	-0.966
(d) SMRTCN(1,1,0.5,2,1,1)	0	-1.760
(e) SMRTCN{2,(0.3,0.7),(0.6,0.5),(5,0),(1,4),(2,1)}	-0.567	-1.714
(f) SMRTCN{2,(0.3,0.7),(0.6,0.5),(0,4),(0.5,4),(0.5,4)}	0	-1.152
(g) SMRTCN{2,(0.8,0.2),(0.5,0.5),(0,0),(0.25,5),(0.25,5)}	0	7.524
(h) Laplace(0,$2^{-1/2}$)	0	3
(i) SMLaplace{2,(0.5,0.5),(0,0),(1,4)}	0	7.671

Table 2.1: The distributions used to generate the scaled errors in the simulation experiment. Let MRTCN($K, \boldsymbol{\pi}_\epsilon, \mathbf{p}, \widetilde{\boldsymbol{\mu}}, \boldsymbol{\sigma}_1^2, \boldsymbol{\sigma}_2^2$) denote a K component mixture of moment restricted two-component normals: $\sum_{k=1}^{K} \pi_{\epsilon k} f_{c\epsilon}(\cdot \mid p_k, \widetilde{\mu}_k, \sigma_{k1}^2, \sigma_{k2}^2)$. Then SMRTN denotes a scaled version of MRTCN, scaled to have variance one. Laplace(μ, b) denotes a Laplace distribution with location μ and scale b. SMLaplace($K, \boldsymbol{\pi}_\epsilon, \mathbf{0}, \mathbf{b}$) denotes a K component mixture of Laplace densities: $\sum_{k=1}^{K} \pi_{\epsilon k}$Laplace$(0, b_k)$, scaled to have variance one. With μ_k denoting the k^{th} order central moments of the scaled errors, the skewness and excess kurtosis of the distribution of scaled errors are measured by the coefficients $\gamma_1 = \mu_3$ and $\gamma_2 = \mu_4 - 3$, respectively. The densities (a)-(f) are light-tailed, whereas the densities (g)-(i) are heavy-tailed. The shapes of these distributions are illustrated in Figure 2.3.

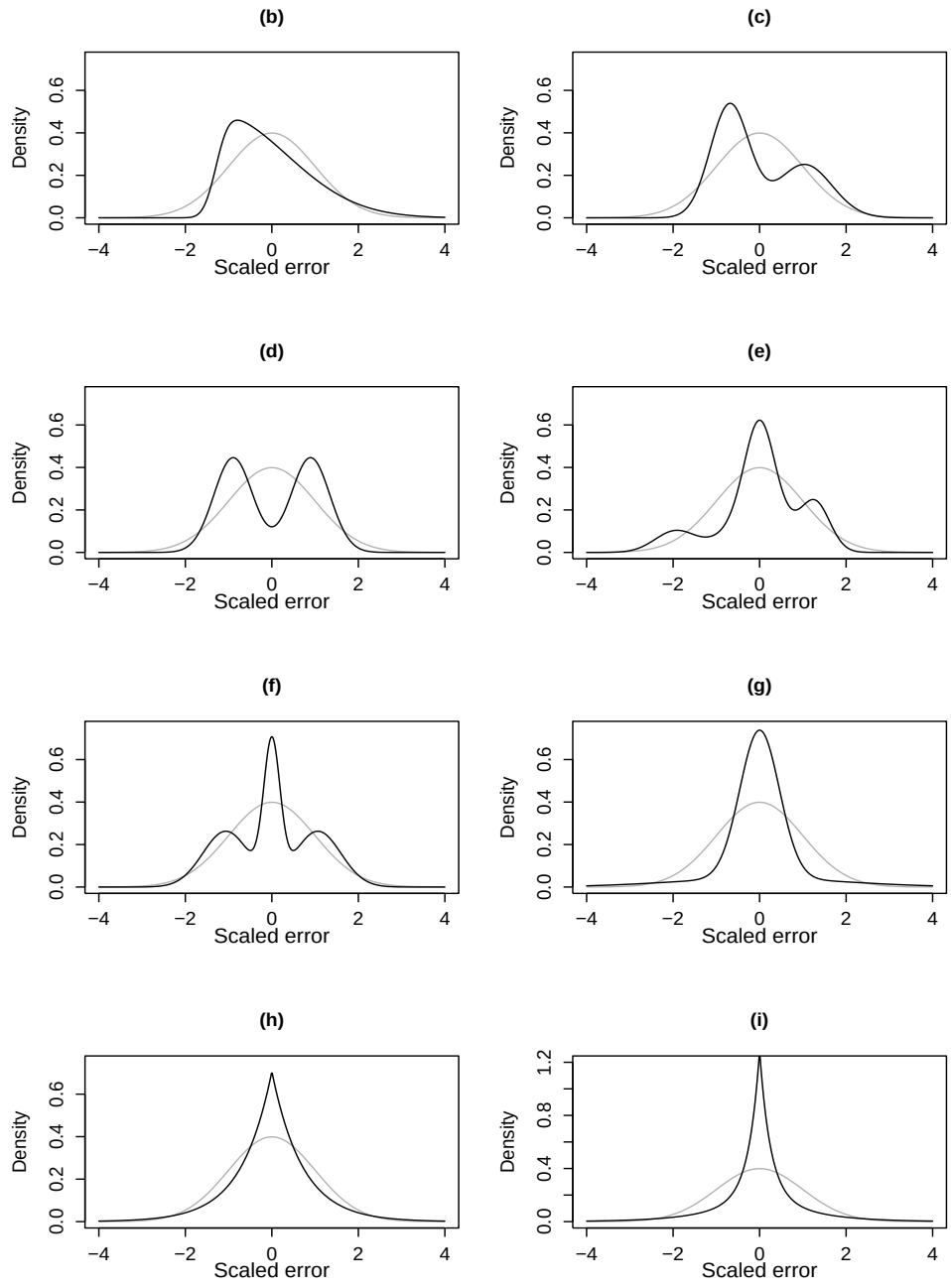

Figure 2.3: The distributions used to generate the scaled errors in the simulation experiment, superimposed over a standard normal density. The different choices cover a wide range of possibilities - (a) standard normal (not shown separately), (b) asymmetric skew-normal, (c) asymmetric bimodal, (d) symmetric bimodal, (e) asymmetric trimodal, (f) symmetric trimodal, (g) symmetric heavy-tailed, (h) symmetric heavy-tailed with a sharp peak at zero and (i) symmetric heavy-tailed with even a sharper peak at zero. The last six cases demonstrate the flexibility of mixtures of moment restricted two-component normals in capturing widely varying shapes.

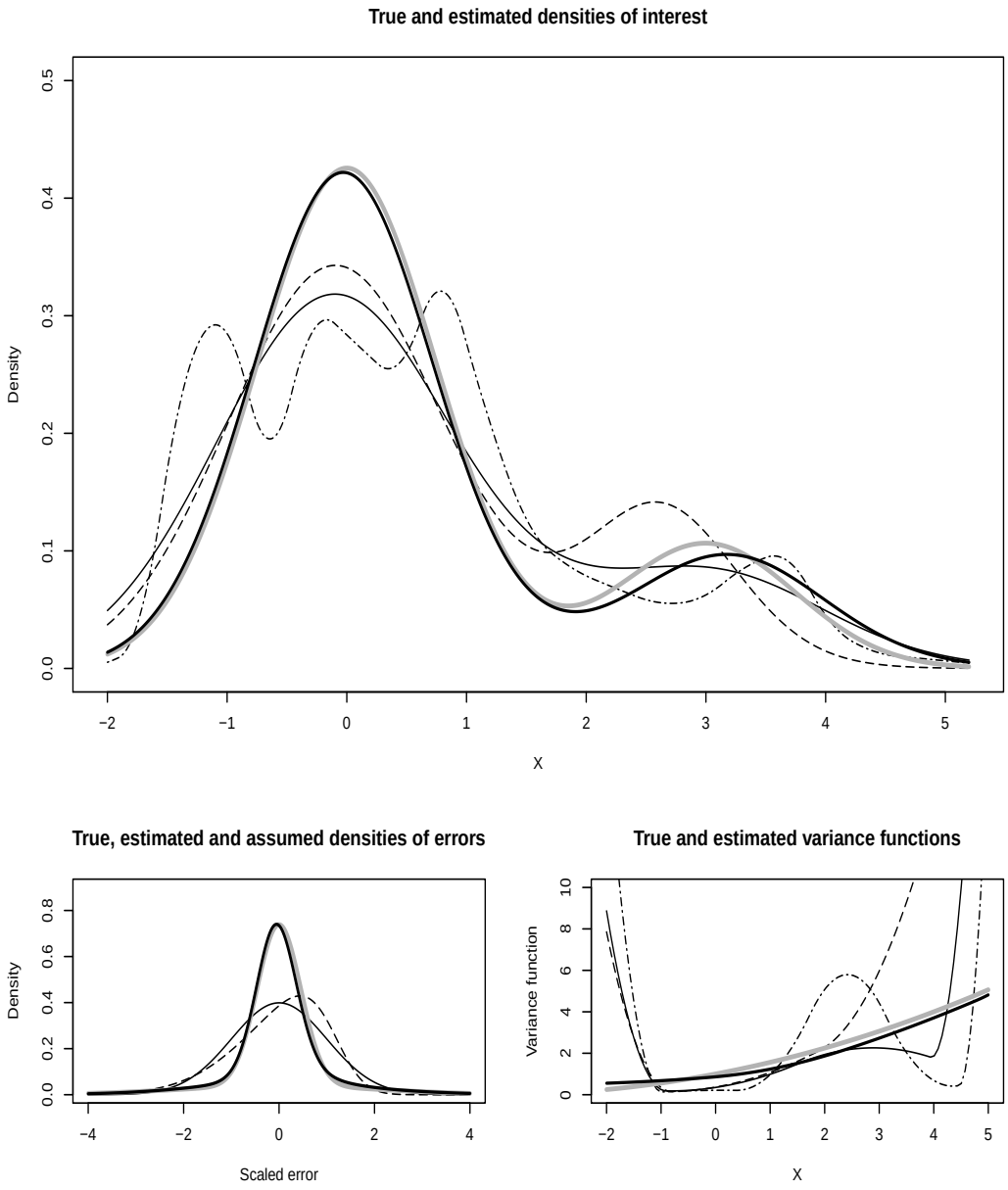

Figure 2.4: Results for heavy-tailed error distribution (g) with sample size n=1000 corresponding to 25^{th} percentile MISEs. The top panel shows the estimated densities under different models. The bottom left panel shows estimated densities of scaled errors under Model-II (dashed line) and Model-III (solid bold line) superimposed over a standard Normal density (solid line). The bottom right panel shows estimated variance functions under different models. For the top panel and the bottom right panel, the solid thin line is for Model-I; the dashed line is for Model-II; the solid bold line is for Model-III; and the dot-dashed line is for the Model of Staudenmayer, et al. (2008). In all three panels the bold gray lines represent the truth.

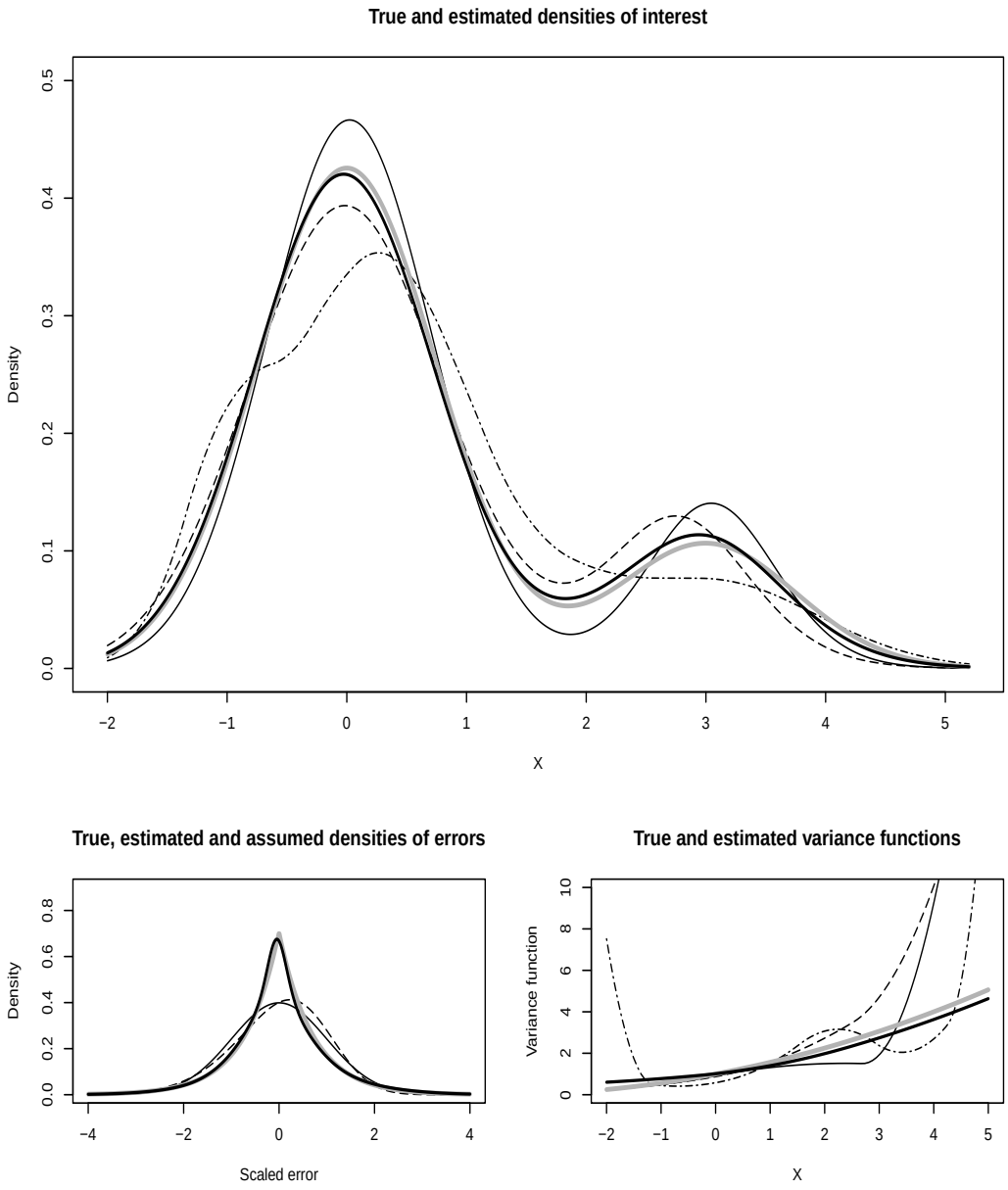

Figure 2.5: Results for heavy-tailed Laplace error distribution (h) with sample size n=1000 corresponding to 25^{th} percentile MISEs. The top panel shows the estimated densities under different models. The bottom left panel shows estimated densities of scaled errors under Model-II (dashed line) and Model-III (solid bold line) superimposed over a standard Normal density (solid line). The bottom right panel shows estimated variance functions under different models. For the top panel and the bottom right panel, the solid thin line is for Model-I; the dashed line is for Model-II; the solid bold line is for Model-III; and the dot-dashed line is for the Model of Staudenmayer, et al. (2008). In all panels the bold gray lines represent the truth.

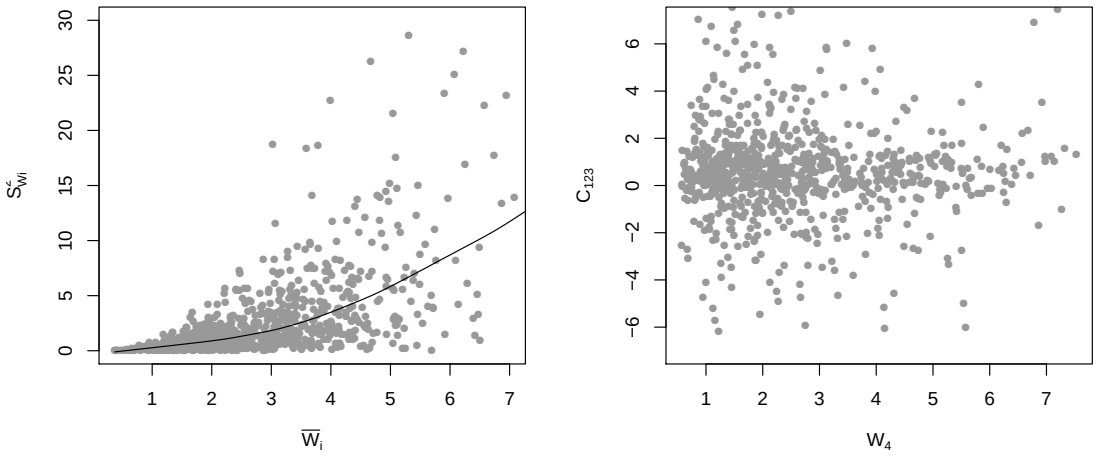

Figure 2.6: Diagnostic plots for reported daily intakes of folate. The left panel shows the plot of $\overline{W}$ vs S^2_W with a simple lowess fit superimposed. The right panel shows the plot of W_4 vs C_{123}.

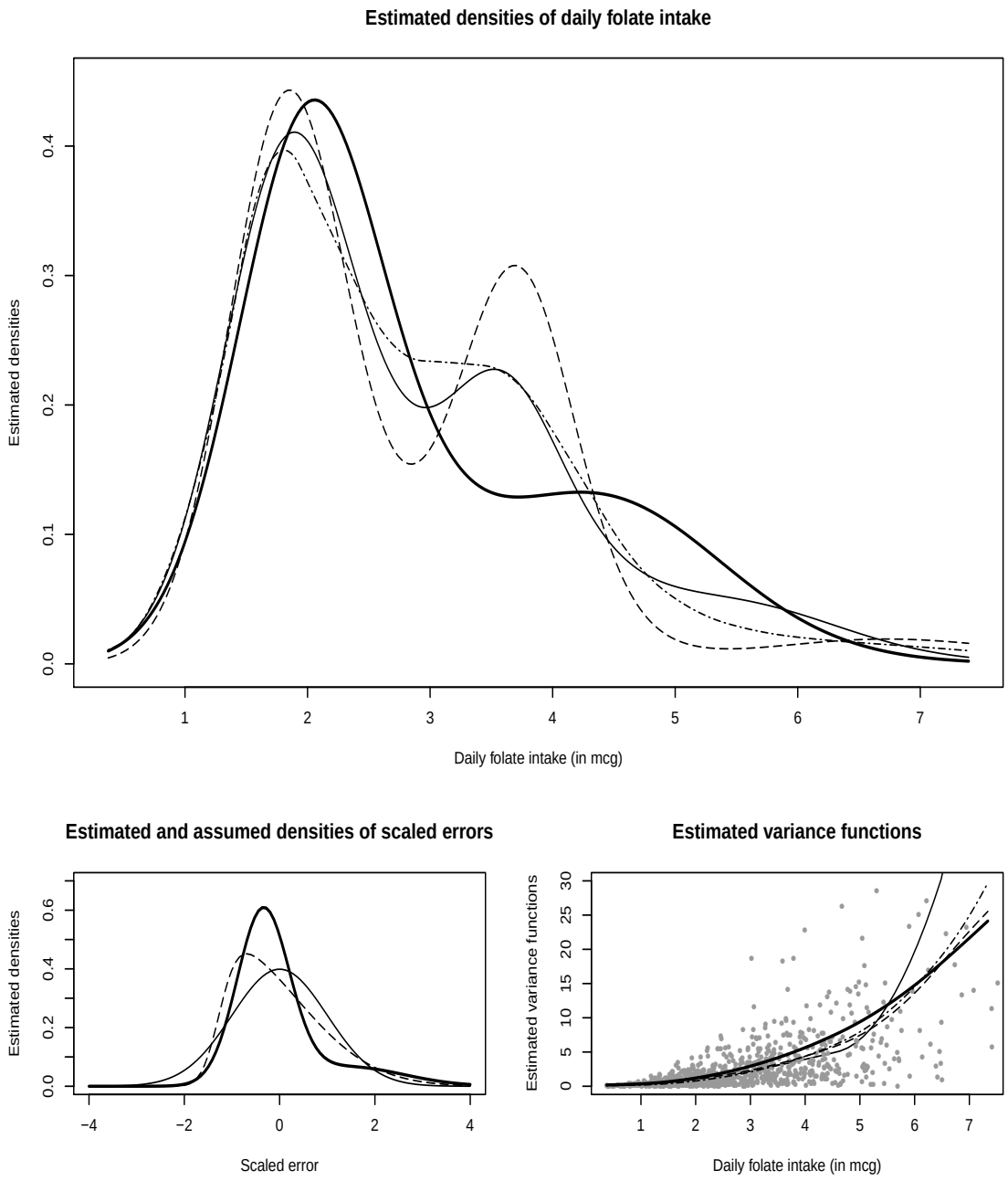

Figure 2.7: Results for data on daily folate intakes from EATS example. The top panel shows the estimated densities of daily folate intake under different models. The bottom left panel shows estimated densities of scaled errors under Model-II (dashed line) and Model-III (solid bold line) superimposed over a standard Normal density (solid line). The bottom right panel shows estimated variance functions under different models. The gray dots represent subject-specific sample means (x-axis) and variances (y-axis). For the top panel and the bottom right panel, the solid thin line is for Model-I; the dashed line is for Model-II; the solid bold line is for Model-III; and the dot-dashed line is for the Model of Staudenmayer, et al. (2008).

True Error Distribution	True X Distribution	Sample Size	MISE $\times 1000$			
			SRB	Model1	Model2	Model3
(a)	50-50 mixture of normals	250	10.15	**5.31**	5.61	5.55
		500	6.64	**3.15**	3.16	3.34
		1000	4.50	**1.96**	2.08	2.21
	80-20 mixture of normals	250	9.60	**4.41**	4.47	4.52
		500	5.30	**2.34**	2.39	2.62
		1000	4.39	**1.31**	1.37	1.39
(b)	50-50 mixture of normals	250	11.79	7.80	**4.41**	4.55
		500	11.85	5.79	**3.11**	3.33
		1000	8.66	4.58	**1.91**	2.21
	80-20 mixture of normals	250	10.74	6.97	**4.52**	4.54
		500	7.94	4.17	**2.27**	2.60
		1000	6.16	3.08	**1.26**	1.39
(c)	50-50 mixture of normals	250	12.61	8.74	5.31	**4.60**
		500	9.27	4.91	3.57	**3.39**
		1000	9.15	4.13	2.53	**1.91**
	80-20 mixture of normals	250	9.27	6.46	4.65	**4.03**
		500	6.67	3.18	2.77	**2.37**
		1000	5.04	2.26	1.40	**1.26**
(d)	50-50 mixture of normals	250	10.10	7.71	9.94	**4.40**
		500	6.54	4.26	7.01	**2.70**
		1000	6.02	3.41	5.58	**1.40**
	80-20 mixture of normals	250	8.18	5.32	5.92	**3.43**
		500	4.45	2.67	4.30	**2.21**
		1000	4.40	1.74	3.31	**1.60**
(e)	50-50 mixture of normals	250	10.03	6.01	5.92	**4.03**
		500	9.38	3.87	3.57	**2.99**
		1000	8.39	2.42	2.25	**1.75**
	80-20 mixture of normals	250	7.82	3.97	4.44	**3.38**
		500	7.62	3.00	2.40	**2.01**
		1000	6.82	1.74	1.45	**1.17**
(f)	50-50 mixture of normals	250	9.35	5.82	6.52	**5.37**
		500	7.18	3.47	3.67	**3.62**
		1000	4.63	2.46	2.62	**2.10**
	80-20 mixture of normals	250	9.17	4.75	4.80	**4.10**
		500	7.35	2.58	2.65	**2.52**
		1000	3.86	1.53	1.60	**1.45**

Table 2.2: Mean integrated squared error (MISE) performance of density deconvolution models described in Section 3.2 of this dissertation (Models I, II and III) compared with the model of Staudenmayer, et al. (2008) (Model SRB) for different light-tailed scaled error distributions. The true variance function was $v(X) = (1 + X/4)^2$. See Section 2.6.1 for additional details. The minimum value in each row is highlighted.

True Error Distribution	True X Distribution	Sample Size	MISE $\times 1000$			
			SRB	Model1	Model2	Model3
(g)	50-50 mixture of normals	250	15.68	11.78	10.38	**3.30**
		500	23.27	15.57	14.85	**2.07**
		1000	49.77	18.91	21.00	**1.12**
	80-20 mixture of normals	250	20.05	8.18	15.99	**3.10**
		500	36.46	10.83	17.23	**1.63**
		1000	48.70	18.53	17.77	**0.92**
(h)	50-50 mixture of normals	250	11.29	6.62	7.01	**5.18**
		500	15.07	8.07	7.24	**3.29**
		1000	18.79	12.04	8.41	**1.99**
	80-20 mixture of normals	250	11.34	7.18	7.05	**4.11**
		500	13.23	7.43	7.53	**2.41**
		1000	22.03	8.64	7.56	**1.25**
(i)	50-50 mixture of normals	250	19.34	7.69	9.90	**3.10**
		500	28.79	17.32	11.02	**2.14**
		1000	54.03	26.78	11.64	**0.94**
	80-20 mixture of normals	250	29.81	16.45	14.76	**2.74**
		500	48.41	20.94	14.99	**1.60**
		1000	57.87	23.80	16.59	**0.83**

Table 2.3: Mean integrated squared error (MISE) performance of density deconvolution models described in Section 3.2 (Models I, II and III) compared with the model of Staudenmayer, et al. (2008) (Model SRB) for different heavy tailed scaled error distributions. The true variance function was $v(X) = (1 + X/4)^2$. See Section 2.6.1 for additional details. The minimum value in each row is highlighted.

True Error Distribution	True X Distribution	Sample Size	MISE $\times 1000$	
			NPM	Model3
(j)	50-50 mixture of normals	250	29.25	**5.25**
		500	23.83	**3.61**
		1000	20.11	**2.45**
	80-20 mixture of normals	250	8.09	**4.62**
		500	6.71	**3.12**
		1000	7.34	**2.05**
(k)	50-50 mixture of normals	250	23.18	**4.81**
		500	20.45	**3.18**
		1000	20.37	**2.13**
	80-20 mixture of normals	250	11.62	**4.42**
		500	8.26	**2.77**
		1000	8.01	**1.43**
(l)	50-50 mixture of normals	250	21.69	**5.65**
		500	17.72	**3.86**
		1000	16.43	**2.67**
	80-20 mixture of normals	250	5.67	**4.71**
		500	3.67	**2.98**
		1000	3.37	**2.01**

Table 2.4: Mean integrated squared error (MISE) performance of Models III compared with the NPM model for different measurement error distributions. See Section 2.6.2 for additional details. The minimum value in each row is highlighted.

	Dietary Component	P-value combined from 4!=24 tests			
		Truncation Limit $\varsigma = 0.05$		Truncation Limit $\varsigma = 0.50$	
		Kendall's τ Test	Spearman's ρ Test	Kendall's τ Test	Spearman's ρ Test
1	Calcium	1	1	0.511	0.984
2	Carbohydrate	1	1	0.824	1
3	Carotene	1	1	0.816	0.993
4	Cholesterol	1	1	0.978	1
5	Copper	1	1	0.982	1
6	Monosaturated Fat	1	1	0.777	1
7	Polysatuared Fat	1	1	1	1
8	Saturated Fat	1	1	0.987	1
9	Fiber	1	1	0.627	0.995
10	Folate	1	1	1	1
11	Iron	1	1	0.996	1
12	Magnesium	1	1	1	1
13	Niacin	1	1	0.910	0.999
14	Phosphorus	0.986	1	0.769	0.986
15	Potassium	1	1	0.989	1
16	Protein	1	1	0.969	1
17	Riboflavin	1	1	1	1
18	Sodium	1	1	0.856	0.999
19	Thiamin	1	1	1	1
20	Vitamin A	1	1	0.999	1
21	Vitamin B6	1	1	0.985	1
22	Vitamin B12	1	1	0.999	1
23	Vitamin C	0.980	1	0.507	0.970
24	Vitamin E	1	1	1	1
25	Zinc	1	1	1	1

Table 2.5: Combined p-values for $4! = 24$ nonparametric tests of association between W_{j_1} and $C_{j_2 j_3 j_4} = \{(W_{j_2} - W_{j_3})/(W_{j_2} - W_{j_4})\}$ for various $j_1 \neq j_2 \neq j_3 \neq j_4$ for 25 regularly consumed dietary components for which daily intakes were recorded in the EATS study. See Section 2.3 for additional details.

3. REGRESSION IN THE PRESENCE OF CONDITIONALLY HETEROSCEDASTIC MEASUREMENT AND REGRESSION ERRORS

3.1 Introduction

In this section we extend the methodology presented in Section 2 to develop a Bayesian semiparametric approach for robust estimation of a regression function when the covariate is measured with error, the density of the covariate, the density of the measurement errors and the density of the regression errors are all unknown, and the variability of both the measurement errors and the regression errors may depend on the associated unobserved value of the covariate through unknown relationships. By 'robust' we mean that we avoid restrictive assumptions common in the literature, such as homoscedasticity and normally distributed measurement and regression errors.

The literature on regression with errors in covariates is extensive. A brief review of the existing literature relevant for our problem is presented here. For a more extensive review of the state of the art see Carroll, et al. (2006) and Buonaccorsi (2010).

The problem of linear regression in the presence of errors in covariates is vast, and besides the references above also includes the classic text by Fuller (1987). More complex problems have also been studied. Cheng and Riu (2006) studied linear models, and considered maximum likelihood, method of moments and generalized least squares estimators for heteroscedastic normally distributed regression and measurement errors. However, they assume that the variances are known and independent of the unobserved value of the covariate. Cook and Stefanski (1994) proposed a simulation-extrapolation (SIMEX) based method that did not make any assumptions about the density of the covariate and the density of the regression errors, but assumes homoscedasticity of both regression and measurement errors. The SIMEX method also required the density of the measurement errors to be known. In the presence of replicated surrogates for the unobserved covariate, Devanarayan and Stefanski (2002) relaxed the homoscedasticity assumptions of the SIMEX approach, but the measurement errors are still required to be normally distributed. Carroll, et al. (1999b) proposed a Bayesian solution to the problem for normally distributed homoscedastic regression and measurement errors. They modeled the unknown density

of the covariate by a finite mixture of normals.

Our focus here is on flexible nonparametric and semiparametric models for all the components. The problem of regression with errors in covariates when the regression and measurement errors are both homoscedastic is studied by Fan and Truong (1993), Carroll, et al. (1999a), Berry, et al. (2002), Carroll and Hall (2004) among others. Fan and Truong (1993) studied deconvoluting kernel type estimators when the density of the measurement errors is known. Carroll, et al. (1999a) studied SIMEX estimators for the nonparametric regression with errors in covariates problem using three different types of models for the regression function, kernel mixtures, smoothing splines, and penalized truncated polynomial splines, but assuming homoscedastic normally distributed measurement errors. Berry, et al. (2002) provided a Bayesian solution to the problem in the presence of normally distributed regression and measurement errors. They also assumed normality of the covariate and modeled the regression function using smoothing splines and penalized mixtures of truncated polynomial splines. Carroll and Hall (2004) considered the problem of estimating a low order estimate of the regression function, rather than the regression function itself. Their method required knowledge of low order moments of the density of the measurement errors that can also be estimated from replicated surrogates. Schennach (2004a, 2004b) studied least squares and Nadaraya-Watson type estimators for nonlinear and nonparametric regression problems, respectively, when measurement error density is unknown but replicated proxies are available and the measurement error in at least one of the replicates is homoscedastic and independent of the covariate. Delaigle and Meister (2007) relaxed the homoscedasticity assumption on the measurement errors but retained it for the regression errors. They developed deconvoluting kernel type estimators for problems when replicated surrogates are available for the unobserved covariates, the density of the regression errors is unknown and homoscedastic, the density of the measurement errors is unknown and heteroscedastic but they are both independent of the covariate.

Conditional heteroscedasticity, as we have seen in Section 2 of this dissertation, can be a very prominent feature of measurement errors in problems of practical importance. The same also remains true for regression errors. This general regression problem has not been addressed in the literature and it is not clear how the general deconvoluting kernel approach or even the automated SIMEX approach can be extended to accommodate conditional heteroscedasticity in regression and measure-

ment errors. On the other hand, Bayesian hierarchical framework can provide a natural way to tackle this otherwise complicated problem by modeling the regression function and the nuisance densities separately through a natural hierarchy. Indeed, a straightforward extension of the methodology developed in Section 2 gives an efficient Bayesian semiparametric solution to the problem. Framed in terms of a Bayesian hierarchical model, we now have to separately model the density of the covariate, the density of the measurement errors, the density of the regression errors and the regression function. The density of the covariate and the density of the measurement errors can be modeled exactly as in Section 2. Recall that in Section 2 we modeled the density of the covariate by a flexible location-scale mixture of normals induced by a Dirichlet process, the measurement errors were factored into 'scaled errors' and 'variance function' components, the density of the scaled errors was modeled using flexible DPMMs, each component of the DPMM being itself a two-component normal mixture with its mean restricted at zero and the variance function was modeled by mixture of B-splines. For reasons discussed in Section 3.2.2 below, modeling the density of the regression errors is actually a harder problem, but we show that in practice the strategy used to model the measurement errors works well for regression errors too. Thus, the only additional function we have to model for the regression problem is essentially the regression function itself and this can be done using flexible mixtures of B-splines in a way very similar to the model for the variance functions.

The section is organized as follows. To make this section somewhat self-contained, in Section 3.2 we revisit the models discussed in Section 2. In this subsection we also describe the model for the regression function. Simulation experiments that compare the performances of our method and the method of Berry, et al. (2002) are presented in Section 3.3, showing that our methods dominate. Section 3.4 presents an application in nutritional epidemiology. Implementation details, such as choice of the hyper-parameters and details of the posterior calculations etc., are moved to Appendix B.

3.2 Models

We consider the problem of robust estimation of the regression relationship between a response Y and a covariate X based on sample in which direct measurements on X are not available, but replicated proxies W for the latent X are available for each sampled unit. Specifically, we assume the data generating model to be

$$\begin{aligned} Y_i &= r(X_i) + U_{Y,i}, \quad i = 1, 2, \ldots, n, && (3.1) \\ W_{ij} &= X_i + U_{W,ij}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m_i \geq 2. && (3.2) \end{aligned}$$

Given X_i, the regression errors $U_{Y,i}$ and the measurement errors $U_{W,ij}$ have mean zero and are conditionally independent. The densities f_X, $f_{U_Y|X}$ and $f_{U_W|X}$ are all unknown. Given r, f_X, $f_{U_Y|X}$ and $f_{U_W|X}$, the likelihood $f_{Y,\mathbf{W}_{1:m}}$ is obtained by the convolution $f_{Y,\mathbf{W}_{1:m}}(Y, \mathbf{W}_{1:m}) = \int f_{U_Y|X}\{Y - r(X)\} \prod_{j=1}^{m} f_{U_{W_j}|X}(W_j - X) f_X(X) dX$. In a Bayesian hierarchical framework, the problem, therefore, reduces to separate problems of modeling the density of the covariate f_X, modeling the conditional densities of the regression and the measurement errors $f_{U_Y|X}$ and $f_{U_W|X}$, and modeling the regression function r.

3.2.1 Density of the Covariate

As in Section 2, we use a Dirichlet process induced mixture of normal kernels, with a conjugate normal-inverse-gamma (NIG) prior on the location and scale parameters

$$\begin{aligned} f_X(X) &= \textstyle\sum_{k=1}^{\infty} \pi_k \, \text{Normal}(X \mid \mu_k, \sigma_k^2), && (3.3) \\ \pi \sim \text{Stick}(\alpha_X), \qquad & (\mu_k, \sigma_k^2) \sim \text{NIG}(\mu_0, \sigma_0^2/\nu_0, \gamma_0, \sigma_0^2). && (3.4) \end{aligned}$$

3.2.2 Conditional Densities of Regression and Measurement Errors

The problem of flexible modeling of conditionally varying measurement error distributions $f_{U_W|X}$ was addressed in Section 2. In the absence of precise covariate information, the problem of modeling the conditional distribution of regression errors $f_{U_Y|X}$ is even harder. First, there are usually multiple proxies but only a single response available for each unknown X. Hence, there is substantially less data available for modeling $f_{U_Y|X}$. Second, the conditional mean of the surrogates, given X, is simply X, so the residuals can be readily calculated. In contrast, to calculate the residuals for regression errors the unknown regression function also needs to be

estimated, and hence the regression residuals are much less informative about the truth.

Fortunately, the methodology introduced in Section 2 for modeling $f_{U_W|X}$ works for $f_{U_Y|X}$ too. The model for $f_{U_W|X}$ described in Section 2 is therefore briefly revisited here. But, because we will essentially be using the same model for both $f_{U_W|X}$ and $f_{U_Y|X}$, to avoid further repetition, the subscripts Y and W are dropped and the generic notation U is used to refer to both U_Y and U_W. In the sections that follow the subscripts reappear as and when necessary. The same convention is followed for different components of the models and the parameters involved. We assume a multiplicative structure for both the measurement errors and the regression errors:

$$U = v^{1/2}(X)\epsilon, \tag{3.5}$$

where the scaled errors ϵ are independently and identically distributed with zero mean and are also independent of X and the variance function v satisfies the identity constraint $v(X_0) = 1$, where X_0 is an interior point of the support of X. The density of the scaled errors f_ϵ and the variance function v can then be modeled exactly as in Section 2. The prior on f_ϵ is a Dirichlet process induced mixture of two component mean restricted mixture of Normals.

$$\begin{aligned} f_\epsilon(\epsilon) &= \textstyle\sum_{k=1}^{\infty} \pi_{\epsilon k}\ f_{c\epsilon}(\epsilon \mid p_k, \widetilde{\mu}_k, \sigma^2_{k1}, \sigma^2_{k2}), \qquad \pi_\epsilon \sim \text{Stick}(\alpha_\epsilon), & (3.6)\\ (p_k, \widetilde{\mu}_k, \sigma^2_{k1}, \sigma^2_{k2}) &\sim \text{Unif}(0,1)\ \text{Normal}(0, \sigma^2_{\widetilde{\mu}})\ \text{IG}(a_\epsilon, b_\epsilon)\ \text{IG}(a_\epsilon, b_\epsilon), & (3.7) \end{aligned}$$

where $f_{c\epsilon}(\epsilon \mid p, \mu_1, \mu_2, \sigma^2_1, \sigma^2_2) = \{p\ \text{Normal}(\epsilon \mid \mu_1, \sigma^2_1) + (1-p)\ \text{Normal}(\epsilon \mid \mu_2, \sigma^2_2)\}$ with $\mu_1 = c_1\widetilde{\mu}, \mu_2 = c_2\widetilde{\mu}$, $c_1 = (1-p)/\{p^2+(1-p)^2\}^{1/2}$ and $c_2 = -p/\{p^2+(1-p)^2\}^{1/2}$. Such a specification, we recall, can capture a large class of mean zero densities on $\mathbb{R}$.

As discussed in Section 2, since $v(X)^{1/2}\epsilon = \{c\ v(X)^{1/2}\}(\epsilon/c)$ for any $c > 0$, the representation of U given by (3.5) is not unique. However, while working on the regression problem we realized that for inference on the regression function r, the variance function v and the density of the scaled errors f_ϵ need not be separately identifiable. Conditional variability of U may simply be obtained as $\text{var}(U \mid X) = v(X)\text{var}(\epsilon)$, and to aid in comparison, versions of f_ϵ adjusted to have unit variance may be retained for each MCMC iteration. The model for v is this specified as a mixture of B-spline basis functions with smoothness inducing priors

on the coefficients as in Section 2, but without any identifiability restriction. That is, we specify

$$\begin{aligned} v(X) &= \textstyle\sum_{j=1}^{J} b_{q,j}(X)\exp(\xi_j) = \mathbf{B}_{q,J}(X)\exp(\boldsymbol{\xi}), && (3.8)\\ p_0(\boldsymbol{\xi} \mid J, \sigma_\xi^2) &\propto \exp\{-\boldsymbol{\xi}^{\mathrm{T}} P \boldsymbol{\xi}/(2\sigma_\xi^2)\}, && (3.9) \end{aligned}$$

where $\boldsymbol{\xi} = \{\xi_1, \xi_2, \ldots, \xi_J\}^{\mathrm{T}}$, $P = D^{\mathrm{T}}D$, where D is a $J \times (J+2)$ matrix such that $D\boldsymbol{\xi}$ computes the second differences in $\boldsymbol{\xi}$; $\mathbf{B}_{q,J} = \{b_{q,1}, b_{q,2}, \ldots, b_{q,J}\}$ is a set of $J = (q+K)$ B-spline bases of degree q defined on an interval $[A, B]$ using knot points $\mathbf{t}_{1:2q+K+1}$.

3.2.3 Regression Function

The problem of flexible modeling of the regression function is now addressed. Specifically, we are interested in models that are numerically stable and lead to easy and efficient posterior computation.

Gaussian processes (Rasmussen and Williams, 2006) are immensely popular and successful for regression problems with precisely measured covariates. For measurement error problems, however, Gaussian process priors are not particularly suitable since the unobserved values of X would be involved in the prior covariance matrix of the regression function and will not be conditionally independent in the posterior, rendering the method computationally complex and numerical unstable. Splines, on the other hand, do not lead to additional complications in the measurement error setup. In regression with errors in covariates, Carroll, et al (1999b) and Berry, et al. (2002) used penalized mixtures of truncated polynomial splines to model the regression function. We model the regression function as a mixture of B-spline bases with a smoothness inducing prior, similar to the model (3.8) for the variance function. In contrast to the model for the variance function v, the exponentiation of the coefficients of the B-spline bases and the imposition of any identifiability constraint are not necessary. A flexible model for the regression function is thus given by

$$\begin{aligned} r(X) &= \textstyle\sum_{j=1}^{J_R} b_{q,j}(X)\xi_{R,j} = \mathbf{B}_{q,J_R}(X)\boldsymbol{\xi}_R, && (3.10)\\ p_0(\boldsymbol{\xi}_R \mid J_R, \sigma_{R,\xi}^2) &\propto \exp\{-\boldsymbol{\xi}_R^{\mathrm{T}} P_R \boldsymbol{\xi}_R/(2\sigma_{R,\xi}^2)\}. && (3.11) \end{aligned}$$

As discussed in Section 2.2.2, the B-splines are locally supported, nearly orthogonal and can be computed using a simple recursion. These unique properties of B-splines make them numerically more stable than polynomial splines. Because B-

splines are also used to model the variance functions, such a model for the regression function also allows reuse of programming codes for fitting different components of the complete hierarchical model.

3.3 Simulation Experiments

Based on M samples $\boldsymbol{\xi}_R^{(m)}$, $m = 1, \ldots, M$, drawn from the posterior, a Monte Carlo estimate $\widehat{r}(X)$ can be obtained as $\widehat{r}(X) = M^{-1} \sum_{m=1}^{M} \mathbf{B}_{q,J_R}(X) \boldsymbol{\xi}_R^{(m)}$. The integrated squared error of estimation of the regression function $r(\cdot)$ by the estimator $\widehat{r}(\cdot)$ is defined as $ISE = \int \{r(X) - \widehat{r}(X)\}^2 dX$. Based on B simulated data sets, a Monte Carlo estimate of the mean integrated squared error (MISE) is given by $MISE_{est} = B^{-1} \sum_{b=1}^{B} \sum_{i=1}^{N} \{r(X_i^{\Delta}) - \widehat{r}^{(b)}(X_i^{\Delta})\}^2 \Delta_i$, where $\{X_i^{\Delta}\}_{i=0}^{N}$ are a set of grid points on the range of X and $\Delta_i = (X_i^{\Delta} - X_{i-1}^{\Delta})$ for all i.

We performed simulation experiments to compare the MISE performance of our method with that of Berry, et al. (2002), referred to as the BCR method henceforth, a naive method, and a deconvoluting kernel based estimator, referred to as the DKE method henceforth. The naive method ignores the measurement errors and treats the subject specific means as the true covariates but accommodates conditional heteroecsdasticity in the regression errors. The DKE method is implemented using the 'DeconNpr' function from the R package 'decon' (Wang and Wang, 2011) allowing subject specific heteroscedasticity. We compared the methods over a wide range of possibilities. The reported estimated MISEs are all based on a grid of 500 equidistant points on $[-2, 2]$ for $B = 200$ simulated data sets. In each case 10,000 MCMC iterations were run and the initial 5,000 iterations were discarded as burn-in. To reduce autocorrelation among the sampled values, the post burn-in samples were thinned by a thinning interval of length 5.

3.3.1 Setup 1: Homoscedasticity and Normally Distributed X

We mimic simulation experiment setups from Berry, et al. (2002). We let $f_X(X) = \text{Normal}(X \mid 0, 1)$, $r(X) = \sin(\pi X/2)/[1 + 2X^2\{\text{sign}(X) + 1\}]$, $U_W \sim \text{Normal}(0, 0.8^2)$, $\text{var}(U_Y) = 0.3^2$ and compare the methods over a factorial combination of (i) two sample sizes $n = 500, 1000$; (ii) two choices for the number of surrogates per subject $m = 2, 3$; and (iii) five different distributions for the regression scaled errors (three light-tailed densities and two heavy-tailed densities, a subset of what we used in Section 2, see Table 3.1 and Figure 3.1).

The results are presented in Table 3.2. The results show that the MISE perfor-

mance of our method is better than the performance of the BCR method in all 20 cases considered, including the case of normally distributed regression errors, when the parametric assumptions of the BCR method are all satisfied. Results produced by our method and the BCR method for this special case are summarized in Figure 3.2. The BCR method uses truncated polynomial splines (P-splines), while we are using B-splines. As opposed to P-splines, B-splines are locally supported and nearly orthogonal, and are therefore numerically more stable than P-splines. This increased numerical stability of our model results in better performance even in situations when the parametric assumption of the BCR model are satisfied. Additional simulation results that support this claim are presented in the Appendix.

3.3.2 Setup 2: Homoscedasticity and Non-Normally Distributed X

Next we keep the error variances constant at $\text{var}(U_Y) = 0.3^2$ and $\text{var}(U_W) = 0.8^2$ and consider the same regression function $r(X) = \sin(\pi X/2)/[1+2X^2\{\text{sign}(X)+1\}]$ as before, but allow all the densities f_X, f_{U_Y} and f_{U_W} to differ from Normality. We now let $f_X(X) = 0.8\ \text{Normal}(X \mid -1, 0.5) + 0.2\ \text{Normal}(X \mid 1, 0.5)$ and compare the methods over a factorial combination of (i) two sample sizes $n = 500, 1000$; (ii) two choices for the number of surrogates per subject $m = 2, 3$; and (iii) five different distributions for the scaled errors (three light-tailed and two heavy-tailed, see Table 3.1 and Figure 3.1). The results are presented in Table 3.3.

3.3.3 Setup 3: Heteroscedasticity and Non-Normally Distributed X

Finally we consider conditionally heteroscedastic errors and let $v_Y(X) = (0.3 + X/8)^2$ and $v_W(X) = (0.8 + X/4)^2$. As before we let $f_X(X) = 0.8\ \text{Normal}(X \mid -1, 0.5) + 0.2\ \text{Normal}(X \mid 1, 0.5)$, $r(X) = \sin(\pi X/2)/[1+2X^2\{\text{sign}(X)+1\}]$ and compare the methods over a factorial combination of (i) two sample sizes $n = 500, 1000$; (ii) two choices for the number of surrogates per subject $m = 2, 3$; (iii) and five different distributions for the scaled errors (three light-tailed and two heavy-tailed, see Table 3.1 and Figure 3.1). The results are presented in Table 3.4. Results for the heavy tailed error distribution (d) are summarized in Figure 3.3.

Results presented in Tables 3.3 and 3.4 show that our method vastly out-performed the BCR model in all 40 cases considered. For example, in Table 3.3, for the symmetric heavy-tailed error distribution (d) with $n = 1000$, the improvement in MISE over the BCR model is $18.17/1.21 \approx 15$ times when there are 2 surrogates per unit and $14.50/0.94 \approx 15$ times when there are 3 surrogates per unit. Similarly, in Ta-

ble 3.4, for the error distribution (d) with $n = 1000$, the improvement in MISE is $23.89/1.49 \approx 16$ times for 2 surrogates per unit and $15.42/1.05 \approx 15$ times for 3 surrogates per unit.

3.3.4 Additional Simulations

The use of B-splines in our model, as opposed to P-splines used in the BCR model, can explain the somewhat surprising results of Section 3.3.1, where our method was shown to outperform the BCR method even when the parametric assumptions of the BCR method were satisfied. Additional simulation experiments that support this claim are presented in the Supplementary Materials, where we compared our method with an improved version of the BCR method, referred to as the BCRB method, that makes the same parametric assumptions as the BCR model but uses B-splines, not P-splines, to model the regression function. We considered two subcases from each of the three scenarios considered above. When the parametric assumptions of the BCR model were true, the BCRB method outperformed our method. In all other cases, our method outperformed the BCRB method.

Additional simulation experiments were also performed to assess the MISE performance of our method when the true error generating densities depart from the multiplicative structural assumption (3.5). Results that suggest our model is fairly robust to such departures are presented in the Supplementary Material.

The results of these additional simulation experiments emphasize the importance of using flexible but numerically stable components for building measurement error models.

3.4 Example

As an illustration of our methodology, we analyze data collected in the Eating at America's Table (EATS) study (Subar, et al., 2001), a large scale epidemiologic study conducted by the National Cancer Institute (NCI) to assess the role of diet in the etiology and prevention of diseases.

The most practical and economical method for collection of dietary data in large epidemiologic studies is the food frequency questionnaire (FFQ). In most studies the respondents receive the FFQs by mail and are instructed to complete the questionnaires independently and return them in postage paid return envelopes. For obvious reasons the data collected by FFQs on dietary intakes typically have a considerable measurement error, and need to be validated prior to or as part of dietary research.

Improved methods can provide a better idea about the relationship between reported FFQs and the true unobserved dietary intakes. The study of the relationship between reported FFQs and the true dietary intakes is, therefore, of great importance in nutritional epidemiology. Other approaches of data collection include 24 hour dietary recalls, where participants are interviewed and their responses recorded by trained professionals. Compared to FFQs, 24 hour recalls are, therefore, much more expensive but the data collected are also more accurate and detailed and can be used to validate the FFQs.

In the EATS study, $n = 965$ participants returned FFQs (Y_i). They were interviewed $m_i = 4$ times over the course of a year and their 24 hour dietary recalls (W_{ij}) were recorded. The true long term average dietary intakes (X_i) are unobserved.

This is a non-standard setting in that Y is not a health outcome, but rather is also a surrogate for X. Ideally one would thus expect both W and Y to be unbiased for X, that is, $E(W \mid X) = E(Y \mid X) = X$. While Y and W are both proxies for X, the 24 hour recalls W_{ij} are recorded by trained personnels after thoroughly conducted interviews, whereas the FFQs Y_i are merely self-reported. As compared to the 24 hour recalls W, the FFQs Y are therefore much less reliable surrogates for the unobserved X, and some departure from the ideal relationship $E(Y \mid X) = X$ may be suspected. Our goal, therefore, is to estimate the relationship between reported FFQs and the true dietary intakes through a flexible regression relationship $E(Y \mid X) = r(X)$, treating the 24 hour recalls W_{ij} as unbiased proxies.

Results for daily intakes of sodium produced by our method and the method by Berry, et al. (2002) (BCR) are summarized in Figure 3.4. Conditional heteroscedasticity of measurements errors is one salient feature of the proxies W_{ij}, clearly identifiable from the plot of subject-specific means versus subject-specific variances. Since Y_i is essentially also a surrogate for X_i, a similar conditional heteroscedasticity pattern is also expected in the errors in the reported FFQs. The BCR method assumes homoscedasticity and normality for the true intakes and regression and measurement errors. Our method, on the other hand, accommodates conditional heteroscedasticity in both regression and measurement errors and also captures departures from normality in their densities and the density of the true intakes, while providing a more robust and realistic estimate of the regression relationship. The results produced by our method indicate that the FFQs are over-reported for low true intakes and are under-reported for high true intakes. The results also indicate that the density of

the true sodium intakes and the densities of the regression and measurement errors are all positively skewed. As expected the estimated conditional heteroscedasticity patterns in the 24 recalls and the FFQs are also very similar. On the other hand, although some departure from the ideal relationship $E(Y \mid X) = X$ is suspected, the regression function estimated by the BCR method is clearly unrealistic. This is not surprising, particularly in view of the strong parametric assumptions made by the BCR method. This example vividly illustrates the importance of the problem we addressed and methodology we described.

3.5 Conclusion

We considered the problem of robust estimation of a regression function in the presence of conditionally heteroscedastic regression and measurement errors. The problem, though extremely important for real world applications, had never been addressed before in the literature. The methodology we described, therefore, makes important contributions to the measurement error literature. Efficiency of the models in estimating the true regression function was illustrated through simulation experiments for a variety of situations. In particular, we showed that our method vastly dominates the method of Berry, et al. (2002). Our method includes normally distributed covariates and homoscedastic normally distributed regression and measurement errors as special cases. In such restricted special scenarios, we also showed that the performance our model is even better than that of Berry, et al. (2002).

Distribution of scaled errors	Skewness (γ_1)	Excess Kurtosis (γ_2)
(a) Normal(0,1)	0	0
(b) SMRTCN(1,1,0.4,2,2,1)	0.499	-0.966
(c) SMRTCN(1,1,0.5,2,1,1)	0	-1.760
(d) SMRTCN{2,(0.8,0.2),(0.5,0.5), (0,0),(0.25,5),(0.25,5)}	0	7.524
(e) Laplace(0,$2^{-1/2}$)	0	3

Table 3.1: The distributions used to generate the scaled errors in the simulation experiments of Section 3.3. SMRTCN$(K, \boldsymbol{\pi}_\epsilon, \mathbf{p}, \widetilde{\boldsymbol{\mu}}, \boldsymbol{\sigma}_1^2, \boldsymbol{\sigma}_2^2)$ denotes the scaled version of a K component mixture of moment restricted two-component normals: $\sum_{k=1}^{K} \pi_{\epsilon k} f_{c\epsilon}(\cdot \mid p_k, \widetilde{\mu}_k, \sigma_{k1}^2, \sigma_{k2}^2)$, scaled to have variance one. Laplace(μ, b) denotes a Laplace distribution with location μ and scale b. With μ_k denoting the k^{th} order central moments of the scaled errors, the skewness and excess kurtosis of the distribution of scaled errors are measured by the coefficients $\gamma_1 = \mu_3$ and $\gamma_2 = \mu_4 - 3$, respectively. The shapes of these densities are illustrated in Figure 3.1.

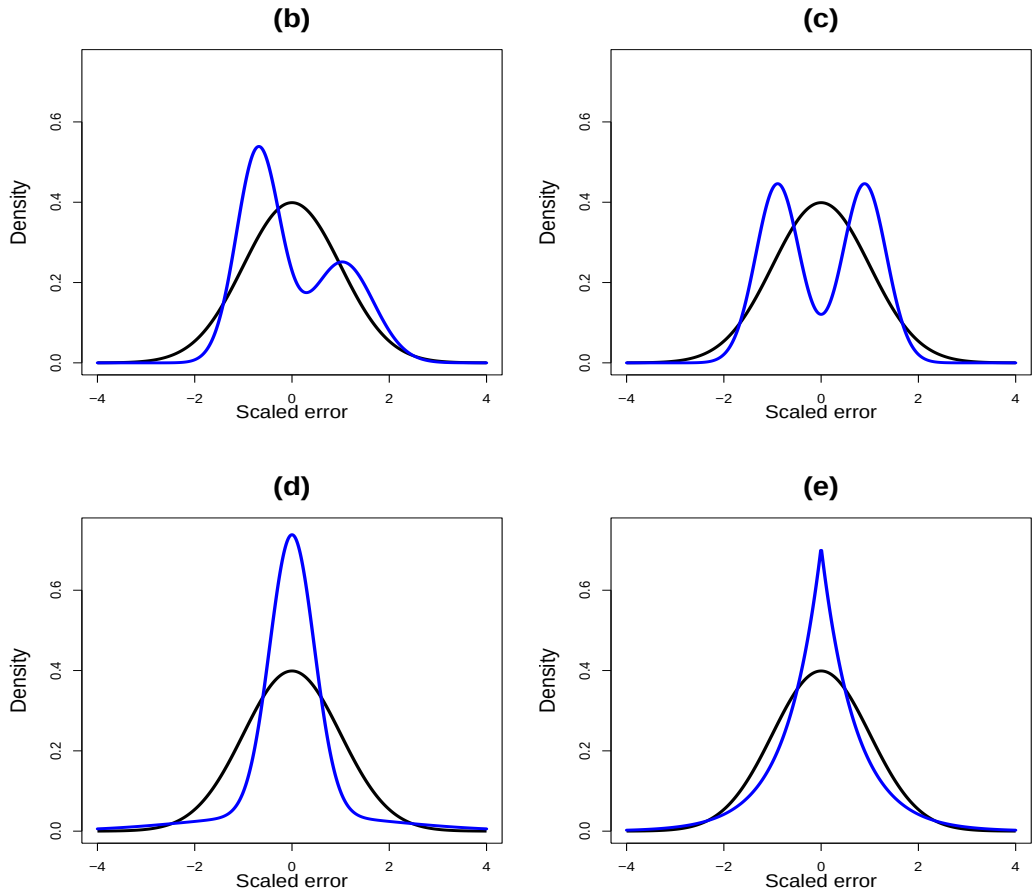

Figure 3.1: The distributions used to generate the scaled regression and measurement errors in simulation experiments, superimposed over a standard normal density - (a) standard normal (not shown separately), (b) asymmetric bimodal, (c) symmetric bimodal, (d) symmetric heavy-tailed and (e) symmetric heavy-tailed with a sharp peak at zero.

True Regression Error Distribution	Sample Size	Number of Replicates	MISE ×100			
			BCR	BSP	Naive	DKE
Normal	500	2	4.98	**2.84**	16.66	24.85
		3	4.09	**1.82**	11.97	22.84
	1000	2	3.11	**1.53**	18.05	20.21
		3	2.42	**0.96**	10.88	16.64
Light-tailed Bimodal Skewed	500	2	4.73	**2.20**	17.75	26.05
		3	4.23	**1.61**	12.29	26.73
	1000	2	3.12	**1.30**	18.45	22.63
		3	2.50	**0.92**	10.80	19.67
Light-tailed Bimodal Symmetric	500	2	4.83	**3.49**	18.60	26.08
		3	4.30	**1.78**	12.50	24.05
	1000	2	3.25	**2.64**	18.60	22.24
		3	2.53	**0.85**	11.13	18.67
Heavy-tailed Symmetric 1	500	2	4.78	**1.82**	17.75	21.69
		3	4.09	**1.38**	11.37	19.22
	1000	2	2.87	**1.10**	19.25	16.42
		3	2.38	**0.76**	11.08	15.90
Heavy-tailed Symmetric 2	500	2	4.77	**2.34**	18.72	24.28
		3	4.14	**1.77**	11.75	23.76
	1000	2	2.99	**1.24**	18.22	17.68
		3	2.41	**0.92**	10.69	16.66

Table 3.2: Mean Integrated Square Error (MISE) performance of our model (BSP) compared to the model of Berry, et al. (2002) (BCR) for homoscedastic simulation experiments in Section 3.3.1, with $X \sim \text{Normal}(0, 1)$, $r(X) = \sin(\pi X/2)/[1 + 2X^2\{\text{sign}(X) + 1\}]$, $U_W \sim \text{Normal}(0, 0.8^2)$ and five different densities for the scaled regression errors (three light-tailed and two heavy-tailed, see Table 3.1 and Figure 3.1 for details) with $\text{var}(U_Y) = 0.3^2$. Our method allows non-normality of X and heteroscedasticity.

True Error Distribution	Sample Size	Number of Replicates	MISE ×100			
			BCR	BSP	Naive	DKE
Normal	500	2	20.30	**6.97**	30.80	43.58
		3	17.29	**3.64**	21.15	37.08
	1000	2	15.85	**3.44**	31.82	37.87
		3	13.18	**2.31**	21.11	32.14
Light-tailed Bimodal Skewed	500	2	19.52	**4.66**	34.44	46.67
		3	16.20	**2.84**	23.86	38.36
	1000	2	14.01	**2.61**	33.57	37.64
		3	11.79	**1.55**	23.22	33.30
Light-tailed Bimodal Symmetric	500	2	20.18	**5.09**	34.67	45.97
		3	17.15	**3.20**	24.08	37.54
	1000	2	15.73	**2.67**	31.61	38.95
		3	13.01	**1.87**	22.52	32.56
Heavy-tailed Symmetric 1	500	2	24.02	**2.19**	20.39	37.08
		3	18.98	**1.76**	16.76	33.49
	1000	2	18.17	**1.21**	21.16	32.00
		3	14.50	**0.94**	17.56	28.96
Heavy-tailed Symmetric 2	500	2	21.74	**4.64**	26.76	40.84
		3	18.25	**3.20**	19.96	37.54
	1000	2	16.99	**2.32**	25.27	33.90
		3	13.48	**1.67**	19.40	29.60

Table 3.3: Mean Integrated Square Error (MISE) performance of our model (BSP) compared to the model of Berry, et al. (2002) (BCR) a naive model that ignores measurement errors (Naive), and a deconvolution kernel estimator (DKE) for the simulation experiments in Section 3.3.2, with $X \sim 0.8\ \text{Normal}(-1, 0.5) + 0.2\ \text{Normal}(1, 0.5)$, $r(X) = \sin(\pi X/2)/[1 + 2X^2\{\text{sign}(X) + 1\}]$ and five different densities for the scaled errors (three light-tailed and two heavy-tailed, see Table 3.1 and Figure 3.1 for details) with $\text{var}(U_Y) = 0.3^2$ and $\text{var}(U_W) = 0.8^2$. Our method allows non-normality of X and heteroscedasticity.

True Error Distribution	Sample Size	Number of Replicates	MISE ×100			
			BCR	BSP	Naive	DKE
Normal	500	2	30.85	**8.47**	21.59	55.00
		3	24.74	**5.21**	15.44	41.57
	1000	2	35.09	**5.80**	18.44	48.03
		3	27.36	**3.93**	11.58	36.14
Light-tailed Bimodal Skewed	500	2	43.92	**7.54**	19.74	52.35
		3	31.47	**2.57**	12.25	39.36
	1000	2	41.53	**2.55**	18.07	46.41
		3	30.10	**1.67**	12.69	34.53
Light-tailed Bimodal Symmetric	500	2	33.72	**4.15**	20.48	57.02
		3	29.08	**2.57**	13.17	40.98
	1000	2	33.50	**2.25**	17.42	51.66
		3	26.84	**1.25**	11.50	35.59
Heavy-tailed Symmetric 1	500	2	26.44	**2.93**	12.84	65.74
		3	15.80	**2.07**	9.56	43.57
	1000	2	23.89	**1.49**	13.40	60.53
		3	15.42	**1.05**	10.05	39.29
Heavy-tailed Symmetric 2	500	2	28.58	**6.11**	16.38	51.92
		3	20.01	**3.89**	11.65	40.99
	1000	2	26.73	**3.44**	15.16	47.53
		3	18.57	**2.31**	10.45	35.83

Table 3.4: Mean Integrated Square Error (MISE) performance of our model (BSP) compared to the model of Berry, et al. (2002) (BCR), a naive model that ignores measurement errors (Naive), and a deconvolution kernel estimator (DKE) for the simulation experiments in Section 3.3.3, with $X \sim 0.8\ \text{Normal}(-1, 0.5) + 0.2\ \text{Normal}(1, 0.5)$, $r(X) = \sin(\pi X/2)/[1 + 2X^2\{\text{sign}(X) + 1\}]$, $v_Y(X) = (0.3 + X/8)^2$, $v_W(X) = (0.8 + X/4)^2$, and five different densities for the scaled errors (three light-tailed and two heavy-tailed, see Table 3.1 and Figure 3.1 for details).

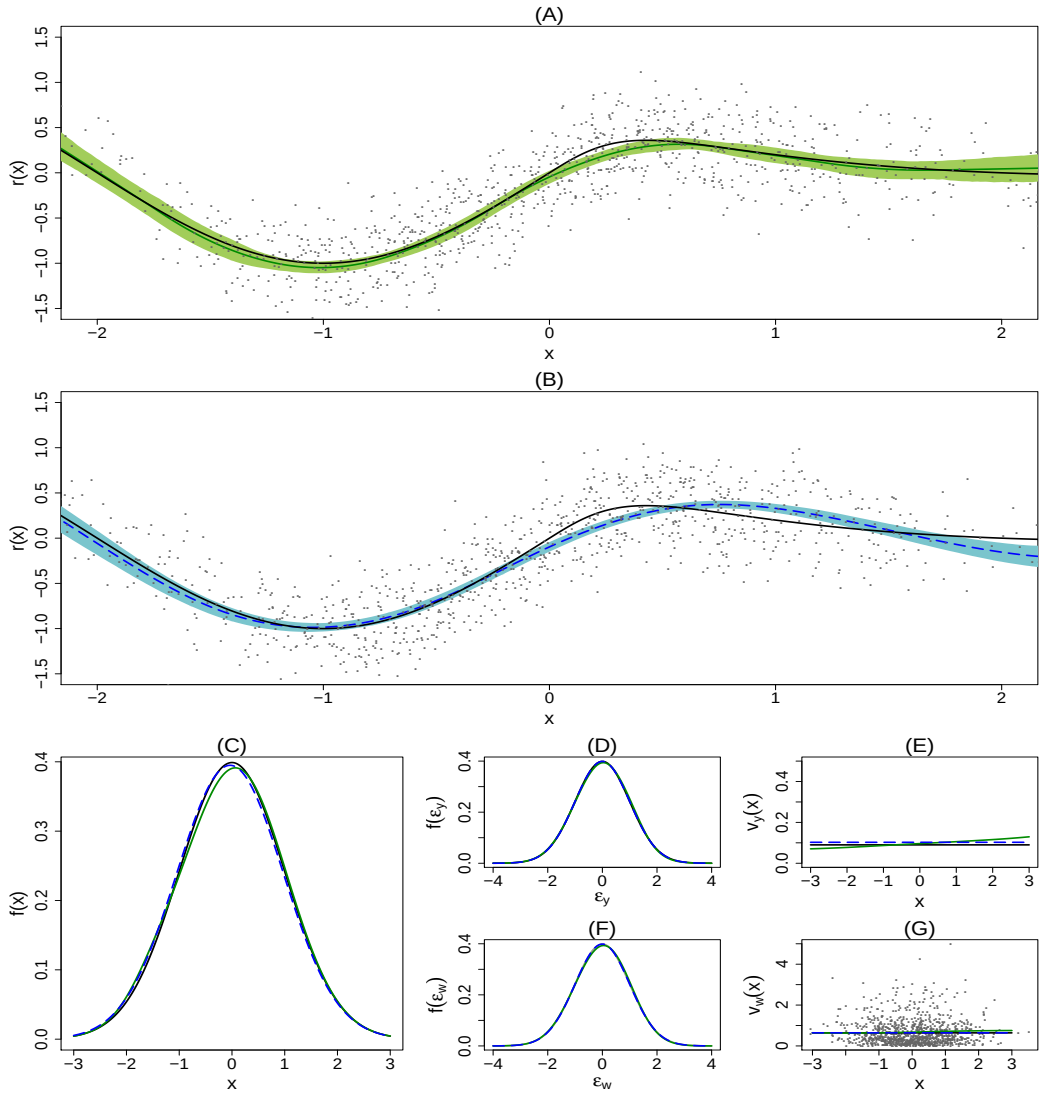

Figure 3.2: Results for our method corresponding to the median MISE in the simulation of Section 3.3.1 when the parametric assumptions of Berry, et al. (2002) are satisfied. Sample size n=1000 and $m = 3$ replicates per subject. In all panels the bold black lines represent the truth, the bold green lines represent the estimates obtained by our method and the dashed blue lines represent the estimates obtained by the method of Berry, et al. (2002) (BCR). (A) The regression function estimated by our method and (B) the regression function estimated by the BCR method. They are presented separately for clarity. In (A) and (B), the gray dots represent estimated posterior mean of the covariate values (x-axis) and the observed responses (y-axis), and the bands represent pointwise 90% credible intervals. (C) The density of the covariate. (D) The density of the scaled regression errors. (E) The variance function of the regression errors. (F) The density of the scaled measurement errors. (G) The variance function of the measurement errors. The gray dots represent subject-specific sample means (x-axis) and variances (y-axis) of the surrogates.

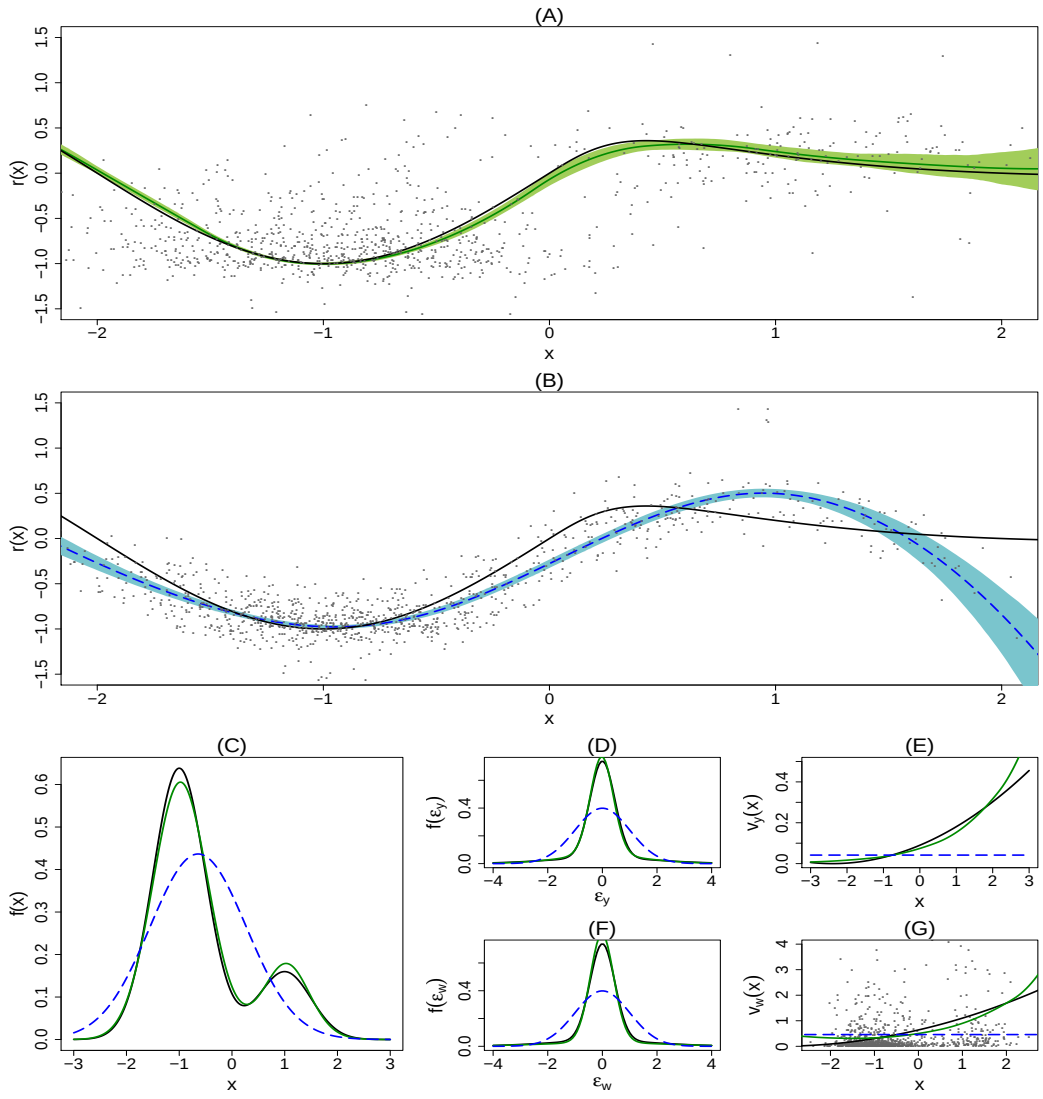

Figure 3.3: Results for heavy-tailed error distribution (d), sample size n=1000 and $m = 3$ replicates per subject corresponding to the median MISEs in the simulation of Section 3.3.3 when X is not Normally distributed, the regression errors and the measurement errors are conditionally heteroscedastic and non-Normal. In all panels the bold black lines represent the truth, the bold green lines represent the estimates obtained by our method and the dashed blue lines represent the estimates obtained by the method of Berry, et al. (2002) (BCR). (A) The regression function estimated by our method and (B) the regression function estimated by the BCR method. They are presented separately for clarity. In (A) and (B), the gray dots represent estimated posterior mean of the covariate values (x-axis) and the observed responses (y-axis), and the bands represent pointwise 90% credible intervals. (C) The density of the covariate. (D) The density of the scaled regression errors. (E) The variance function of the regression errors. (F) The density of the scaled measurement errors. (G) The variance function of the measurement errors. The gray dots represent subject-specific sample means (x-axis) and variances (y-axis) of the surrogates.

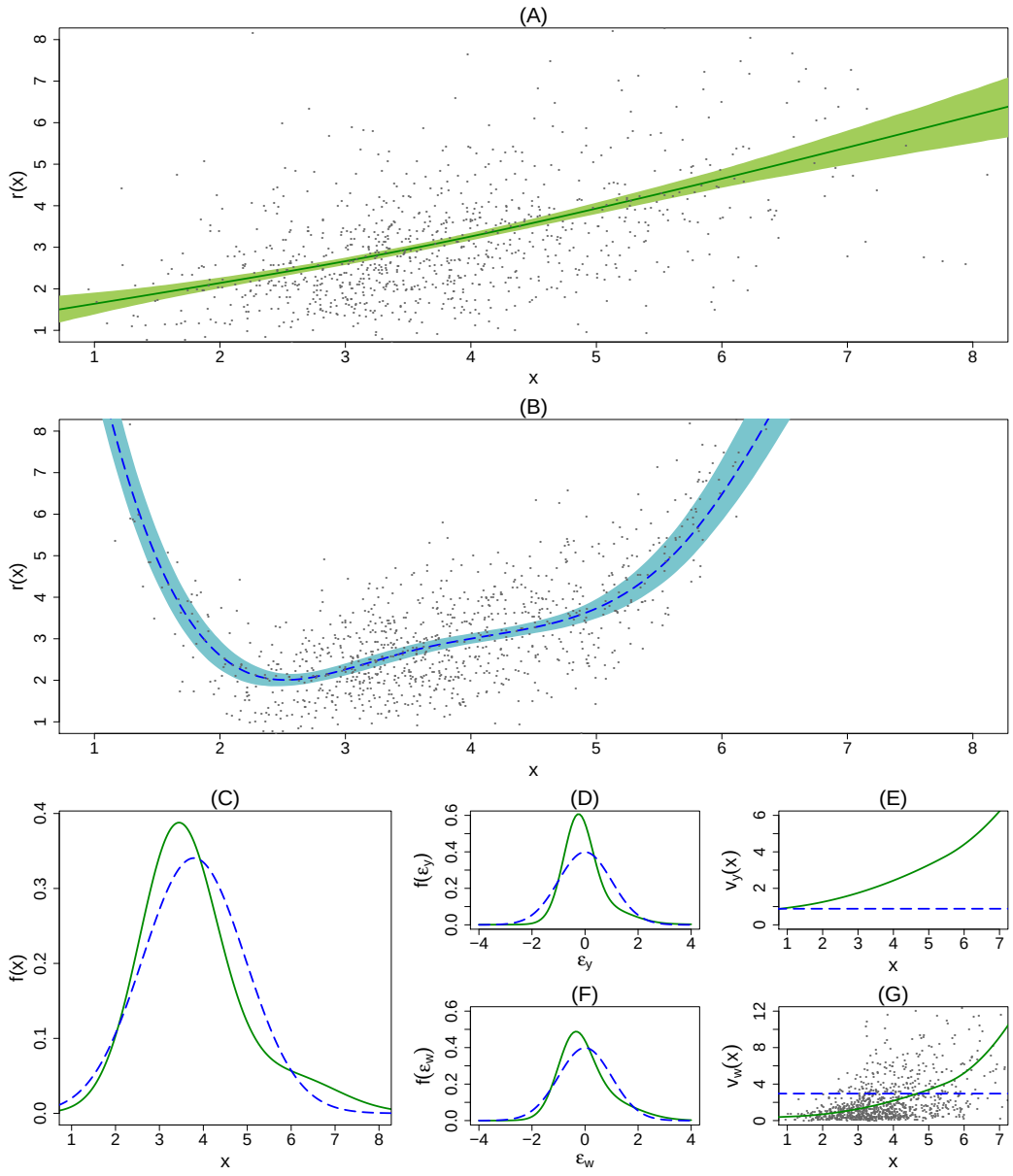

Figure 3.4: Results for sodium from the EATS data set. In all panels the bold green lines represent the estimates obtained by our method and the blue dashed lines represent the estimates obtained by the method of Berry, et al. (2002). (A) The regression function estimated by our method and (B) the regression function estimated by the BCR method. They are presented separately for clarity. In (A) and (B), the gray dots represent estimated posterior mean of the covariate values (x-axis) and the observed responses (y-axis), and the bands represent point wise 90% credible intervals. (C) The density of the covariate. (D) The density of the scaled regression errors. (E) The variance function of the regression errors. (F) The density of the scaled measurement errors. (G) The variance function of the measurement errors. The gray dots represent subject-specific sample means (x-axis) and variances (y-axis) of the surrogates.

4. MULTIVARIATE DENSITY DECONVOLUTION IN THE PRESENCE OF MEASUREMENT ERRORS

4.1 Introduction

The problem of univariate density deconvolution was discussed in Section 2 of this dissertation. In this section we take up the multivariate problem.

In sharp contrast to the univariate case, the literature on multivariate density deconvolution is quite sparse. We can only mention Masry (1991), Youndjé and Wells (2008), Comte and Lacour (2013) and Bovy, et al. (2011). The first three considered deconvoluting kernel based approaches assuming the measurement errors to be distributed independently from the vector of interest density according to a completely known probability law. Bovy, et al. (2011) modeled the density of interest using flexible mixtures of multivariate normal kernels and assumed the measurement errors to be distributed according to multivariate normal probability laws with known covariance matrices, independently from the variable of interest. As in the case of univariate problems, the assumptions of fully specified measurement error distribution, known covariance matrices, and independence from the variables of interest are highly restrictive for most practical applications.

The focus of this Section is on multivariate density deconvolution when the distribution of the measurement errors is not known but replicated proxies are available for each subject. We consider two types of scenarios - (a) when the measurement errors are independently distributed from the vector of interest and (b) when the variability of different components of the measurement errors depends on the associated unobserved value of the vector of interest through unknown relationships. The latter problem is important again in nutritional epidemiology where nutritionists are typically interested not just in the consumption behaviors of individual dietary components but also in their joint consumption patterns, and, as we have seen in Section 2, the data, available in the form of dietary recalls, are contaminated by measurement errors that show strong patterns of conditional heteroscedasticity.

As in Section 2, we use mixture models to estimate both the density of interest and the density of the measurement errors but the multivariate nature of the problem brings in new modeling challenges and computational obstacles that preclude straightforward extension of the univariate deconvolution approaches developed in

Section 2. Instead of using infinite mixtures induced by Dirichlet processes, we now use finite mixtures of multivariate normal kernels with symmetric Dirichlet priors on the mixture probabilities. The use of finite mixtures and exchangeable priors greatly reduces computational complexity while retaining essentially the same flexibility as that of their infinite dimensional counterparts.

Additionally, we also exploit the exchangeability of symmetric Dirichlet priors and some basic properties of multivariate normal distributions and finite mixture models to come up with a novel strategy that enables us to enforce a required zero mean restriction on the measurement errors. The technique proposed in this section, as opposed to the one adopted in Section 2, is particularly suitable for multivariate problems.

It is well known that inverse Wishart priors, due to their dense parametrization, are not suitable for modeling covariance matrices in high dimensional applications. In multivariate deconvolution problems the issue is further complicated since the vector of interest and the measurement errors, the two vector valued variables whose densities we need to model, are both latent. This results in numerically unstable estimates even for small dimensions, particularly when the true covariance matrices are sparse and the likelihood function is of complicated form. To reduce the effective number of parameters required to be estimated, we consider factor-analytic representation of the component specific covariance matrices with sparsity inducing shrinkage priors on the factor loading matrices.

In Section 2, our models for conditionally heteroscedastic univariate measurement errors were somewhat inspired by models for conditionally heteroscedastic regression errors available in the existing literature (Pati and Dunson, 2013; Pelenis, 2014). In contrast, as we have detailed in Appendix C, existing covariance regression techniques are not quite relevant for modeling conditionally heteroscedastic multivariate measurement errors. The models that we formulate cater to these somewhat unique properties of conditionally heteroscedastic multivariate measurement errors but result in a complicated likelihood function that again gives rise to significant computational hurdles. We overcome these obstacles by designing a novel two-stage procedure that exploits the same properties of conditionally heteroscedastic multivariate measurement errors to our advantage. The procedure first estimates the variance functions accommodating conditional heteroscedasticity in the measurement errors using reparametrized versions of the corresponding univariate submodels. The es-

timates obtained in the first stage are then plugged-in to estimate the remaining parameters in the second stage. Having two separate estimation stages, in some sense, the multivariate deconvolution models for conditionally heteroscedastic measurement errors are not purely Bayesian. But they show good empirical performance and, with no other solution available in the existing literature, they provide at least workable starting points towards more sophisticated methodology.

For a review of finite mixture models and mixtures of latent factor analyzers, without moment restrictions or sparsity inducing priors and with applications in measurement error free scenarios, see Fokoué and Titterington (2003), Frühwirth-Schnatter (2006), Mengersen, et al. (2011) and the references therein.

The Section is organized as follows. Section 4.2 details the models. Section 4.5 discusses the choice of hyper-parameters. Section 4.6 describes Markov chain Monte Carlo (MCMC) algorithms to sample from the posterior. The two-step procedure to estimate the variance functions for conditionally heteroscedastic measurement errors is detailed in Section 4.7. Section 4.8 presents simulation studies comparing the proposed deconvolution methods to a naive method that ignores measurement errors. An application of the proposed methodology in estimation of the joint consumption pattern of dietary intakes from contaminated 24 hour recalls collected in the EATS study is presented in Section 4.9. Section 4.10 presents concluding remarks.

We present our detailed arguments in favor of finite mixture models for the multivariate problem in Appendix C, pointing out, in particular, the close connections and the subtle differences the adopted finite dimensional approaches have with the infinite dimensional Dirichlet process based approaches and explaining in detail how these properties are exploited to achieve significant reduction in computational complexity while retaining the major advantages of infinite dimensional mixture models including model flexibility and automated model selection and model averaging. Appendix C also collects additional figures. Theoretical results showing flexibility of the proposed models are presented in Appendix D.

4.2 Deconvolution Models

The goal is to estimate the unknown joint density of a p-dimensional multivariate random variable $\mathbf{X}$. There are $i = 1, 2, \ldots, n$ subjects. Precise measurements of $\mathbf{X}$ are not available. Instead, for $j = 1, 2, \ldots, m_i$, replicated proxies $\mathbf{W}_{ij}$ contaminated with measurement errors $\mathbf{U}_{ij}$ are available for each subject. The replicates are assumed to be generated by the model

$$\mathbf{W}_{ij} \;=\; \mathbf{X}_i + \mathbf{U}_{ij}, \tag{4.1}$$

where $\mathbf{X}_i$ is the unobserved true value of $\mathbf{X}$; given $\mathbf{X}_i$, $\mathbf{U}_{ij}$ are independently distributed with $E(\mathbf{U}_{ij} \mid \mathbf{X}_i) = \mathbf{0}$. The density of $\mathbf{X}$ is denoted by $f_{\mathbf{X}}$. The implied conditional distributions of $\mathbf{W}_{ij}$ and $\mathbf{U}_{ij}$, given $\mathbf{X}_i$, is denoted by $f_{\mathbf{W}|\mathbf{X}}$ and $f_{\mathbf{U}|\mathbf{X}}$, respectively. The marginal density of $\mathbf{W}_{ij}$ is denoted by $f_{\mathbf{W}}$.

4.3 Modeling the Density of Interest

We model $f_{\mathbf{X}}$ as a mixture of multivariate normal kernels

$$f_{\mathbf{X}}(\mathbf{X}) = \textstyle\sum_{k=1}^{K_{\mathbf{X}}} \pi_{\mathbf{X},k}\ \mathrm{MVN}_p(\mathbf{X} \mid \boldsymbol{\mu}_{\mathbf{X},k}, \boldsymbol{\Sigma}_{\mathbf{X},k}), \tag{4.2}$$

where $\mathrm{MVN}_p(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a p-dimensional multivariate normal density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For the rest of this section, the suffix $\mathbf{X}$ is kept implicit to keep the notation clean. Given K, we assign conjugate priors to the mixture probability vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^{\mathrm{T}}$, the component specific means $\boldsymbol{\mu}_k$ and the component specific covariance matrices $\boldsymbol{\Sigma}_k$:

$$\boldsymbol{\pi} \;\sim\; \mathrm{Dir}(\alpha/K, \ldots, \alpha/K), \tag{4.3}$$

$$\boldsymbol{\mu}_k \;\sim\; \mathrm{MVN}_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \quad \boldsymbol{\Sigma}_k \sim \mathrm{IW}_p(\nu_0, \boldsymbol{\Psi}_0). \tag{4.4}$$

Here $\mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$ denotes a finite dimensional Dirichlet distribution on the K-dimensional unit simplex with concentration parameter $(\alpha_1, \ldots, \alpha_K)$, and $\mathrm{IW}_p(\nu, \boldsymbol{\Psi})$ denotes an inverse Wishart density on the space of $p \times p$ positive definite matrices with mean $\boldsymbol{\Psi}/(\nu - p - 1)$. The symmetry of the assumed Dirichlet prior helps in reducing computational complexity. Provided K is sufficiently large, a carefully chosen α can impart the posterior with certain properties that simplify model selection and model averaging issues. Detailed discussions of these topics are deferred to Section 4.5 and

Appendix C.1.

In complex high dimensional problems, the dense parameterization of inverse Wishart prior may result in numerically unstable estimates, particularly when the covariance matrices are sparse. In a deconvolution problem the issue is compounded further by the nonavailability of the true $\mathbf{X}_i$'s. Unlike an ordinary density estimation problem, the density of interest $f_{\mathbf{X}}$ is not the data generating density and only the proxies $\mathbf{W}_{ij}$ contaminated with measurement errors $\mathbf{U}_{ij}$, that are often highly variable, are available for model estimation. To reduce the effective number of parameters to be estimated, we consider a parsimonious factor-analytic representation of the component specific covariance matrices:

$$\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_k \boldsymbol{\Lambda}_k^{\mathrm{T}} + \boldsymbol{\Omega}, \tag{4.5}$$

where $\boldsymbol{\Lambda}_k$ are $p \times q_k$ factor loading matrices and $\boldsymbol{\Omega}$ is a diagonal matrix with non-negative entries. In practical applications q_k will typically be much smaller than p, inducing sparse characterizations of the unknown covariance matrices $\boldsymbol{\Sigma}_k$. Letting C_i denote the cluster label associated with each $\mathbf{X}_i$, model (4.2) can then be equivalently represented as

$$\Pr(C_i = k) = \pi_k, \tag{4.6}$$

$$(\mathbf{X}_i \mid C_i = k) = \boldsymbol{\mu}_k + \boldsymbol{\Lambda}_k \boldsymbol{\eta}_i + \boldsymbol{\Delta}_i, \tag{4.7}$$

$$\boldsymbol{\eta}_i \sim \mathrm{MVN}_p(\mathbf{0}, \mathbf{I}_p), \qquad \boldsymbol{\Delta}_i \sim \mathrm{MVN}_p(\mathbf{0}, \boldsymbol{\Omega}), \tag{4.8}$$

where $\boldsymbol{\eta}_i$ are latent factors and $\boldsymbol{\Delta}_i$ are errors with covariance $\boldsymbol{\Omega} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$.

The above characterization of $\boldsymbol{\Sigma}_k$ is not unique, since for any semi-orthogonal matrix $\mathbf{P}$ the loading matrix $\boldsymbol{\Lambda}_k^1 = \boldsymbol{\Lambda}_k \mathbf{P}$ also satisfies (4.5). However, in a Bayesian framework, when interest lies primarily in estimating the density $f_{\mathbf{X}}$, identifiability of the latent factors is not required. This also allows the loading matrices to have a-priori potentially infinite number of columns. Sparsity inducing priors, that favor more shrinkage as the column index increases, can then be used to shrink the redundant columns towards zero. In this dissertation, we do this by adapting the shrinkage prior proposed in Bhattacharya and Dunson (2011) that allows easy posterior computation. Let $\boldsymbol{\Lambda}_k = ((\lambda_{k,jh}))_{j=1,h=1}^{p,\infty}$, where j and h denote the row and the

column indices, respectively. For $\ell = 1, \ldots, \infty$, we assign

$$\lambda_{k,jh} \sim \text{Normal}(0, \phi_{k,jh}^{-1}\tau_{k,h}^{-1}), \qquad \phi_{k,jh} \sim \text{Ga}(\nu/2, \nu/2), \tag{4.9}$$

$$\tau_{k,h} \sim \prod_{\ell=1}^{h} \delta_{k,\ell}, \qquad \delta_{k,\ell} \sim \text{Ga}(a_\ell, 1), \qquad \sigma_j^2 \sim \text{Inv-Ga}(a_\sigma, b_\sigma). \tag{4.10}$$

Here $\text{Ga}(\alpha, \beta)$ denotes a Gamma distribution with shape parameter α and rate parameter β and $\text{IG}(a, b)$ denotes an inverse-Gamma distribution with shape parameter a and scale parameter b. In the k^{th} component factor loading matrix $\mathbf{\Lambda}_k$, the parameters $\{\phi_{k,jh}\}_{j=1}^{p}$ control the local shrinkage of the elements in the h^{th} column, whereas $\tau_{k,h}$ controls the global shrinkage. When $a_h > 1$ for $h = 2, \ldots, \infty$, the sequence $\{\tau_{k,h}\}_{h=1}^{\infty}$ becomes stochastically increasing and thus favors more shrinkage as the column index h increases.

In addition to inducing adaptive sparsity and hence imparting the model with numerical stability, by favoring more shrinkage as the column index increases, the shrinkage priors play another important role in making the proposed factor analytic model highly robust to misspecification of the number of latent factors, allowing us to adopt simple strategies to determine the number of latent factors to be included in the model in practice. Details are deferred to Section 4.5.

Throughout the rest of this dissertation, mixtures with inverse Wishart prior on the covariance matrices will be referred to as MIW models and mixtures of latent factor analyzers will be referred to as MLFA models.

4.4 Modeling the Density of the Measurement Errors

4.4.1 Independently Distributed Measurement Errors

In this section, we develop models for measurement errors assuming independence from the variables of interest. This remains the most extensively researched deconvolution problem for both univariate and multivariate cases. The techniques developed in this section will also provide crucial building blocks for more realistic models in Section 4.4.2. The measurement errors and the density of the measurement errors are now denoted by $\boldsymbol{\epsilon}_{ij}$ and $f_{\boldsymbol{\epsilon}}$, respectively, for reasons to become obvious shortly in Section 4.4.2.

As in Section 4.3, a mixture of multivariate normals can be used to model the

density $f_{\boldsymbol{\epsilon}}$ but the model now has to satisfy a mean zero constraint. That is

$$f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) = \sum_{k=1}^{K_{\boldsymbol{\epsilon}}} \pi_{\boldsymbol{\epsilon},k} \, \mathrm{MVN}_p(\boldsymbol{\epsilon} \mid \boldsymbol{\mu}_{\boldsymbol{\epsilon},k}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon},k}), \tag{4.11}$$

$$\text{subject to } \sum_{k=1}^{K_{\boldsymbol{\epsilon}}} \pi_{\boldsymbol{\epsilon},k} \boldsymbol{\mu}_{\boldsymbol{\epsilon},k} = \mathbf{0}. \tag{4.12}$$

To get numerically stable estimates of the density of the errors, latent factor characterization of the covariance matrices with sparsity inducing shrinkage priors as in Section 4.3 may again be used. Details are curtailed to avoid unnecessary repetition and we only present the mechanism to enforce the zero mean restriction on the model. The suffix $\boldsymbol{\epsilon}$ is again dropped in favor of cleaner notation. In later sections, the suffixes $\mathbf{X}$ and $\boldsymbol{\epsilon}$ reappear to distinguish between the parameters associated with $f_{\mathbf{X}}$ and $f_{\boldsymbol{\epsilon}}$, when necessary.

Without the mean restriction and under conjugate multivariate normal priors $\boldsymbol{\mu}_k \sim \mathrm{MVN}_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, the posterior full conditional of $\boldsymbol{\mu}^{Kp\times 1} = (\boldsymbol{\mu}_1^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_K^{\mathrm{T}})^{\mathrm{T}}$ is given by

$$\mathrm{MVN}_{Kp}\left\{ \begin{pmatrix} \boldsymbol{\mu}_1^0 \\ \boldsymbol{\mu}_2^0 \\ \vdots \\ \boldsymbol{\mu}_K^0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_1^0 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2^0 & \ldots & \mathbf{0} \\ \vdots & & & \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\Sigma}_K^0 \end{pmatrix} \right\} \equiv \mathrm{MVN}_{Kp}(\boldsymbol{\mu}^0, \boldsymbol{\Sigma}^0), \tag{4.13}$$

where $\boldsymbol{\epsilon}_{ij}$ and other conditioning variables are implicitly understood. Explicit expressions of $\boldsymbol{\mu}^0$ and $\boldsymbol{\Sigma}^0$ in terms of the conditioning variables can be found in Section 4.6. The posterior full conditional of $\boldsymbol{\mu}$ under the mean restriction can then be obtained easily by further conditioning the distribution in (4.13) by $\boldsymbol{\mu}_R = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k = 0$ and is given by

$$(\boldsymbol{\mu} \mid \boldsymbol{\mu}_R = \mathbf{0}) \sim \mathrm{MVN}_{Kp}(\boldsymbol{\mu}^0 - \boldsymbol{\Sigma}_{1,R}^0 \boldsymbol{\Sigma}_{R,R}^{0-1} \boldsymbol{\mu}_R^0, \boldsymbol{\Sigma}^0 - \boldsymbol{\Sigma}_{1,R}^0 \boldsymbol{\Sigma}_{R,R}^{0-1} \boldsymbol{\Sigma}_{R,1}^0), \tag{4.14}$$

where $\boldsymbol{\mu}_R^0 = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k^0 = E(\boldsymbol{\mu}_R)$, $\boldsymbol{\Sigma}_{k,K} = \pi_k \boldsymbol{\Sigma}_k^0 = \mathrm{cov}(\boldsymbol{\mu}_k, \boldsymbol{\mu}_R)$, $\boldsymbol{\Sigma}_{R,R}^0 = \boldsymbol{\Sigma}_{K+1,K+1} = \sum_{k=1}^{K} \pi_k^2 \boldsymbol{\Sigma}_k^0 = \mathrm{cov}(\boldsymbol{\mu}_R)$, and $\boldsymbol{\Sigma}_{R,1}^0 = (\boldsymbol{\Sigma}_{1,K+1}, \boldsymbol{\Sigma}_{2,K+1}, \ldots, \boldsymbol{\Sigma}_{K,K+1})$. To sample from the singular density given in (4.13), we can first sample from the non-singular distribution of $\{(\boldsymbol{\mu}_1^{\mathrm{T}}, \boldsymbol{\mu}_2^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_{K-1}^{\mathrm{T}})^{\mathrm{T}} \mid \boldsymbol{\mu}_R = \mathbf{0}\}$, which can also be trivially obtained from (4.14), and then set $\boldsymbol{\mu}_K = -\sum_{k=1}^{K-1} \pi_k \boldsymbol{\mu}_k / \pi_K$.

Two remarks are in order. First, note that the symmetric Dirichlet prior on the

mixture probabilities plays an additional implicit but important role here. Although we have used the K^{th} component to enforce the mean restriction, under exchangeable Dirichlet priors the posterior is also invariant to permutations of the mixture labels making all the components equally deserving candidates for this fix and the specific choice of any particular component irrelevant. Second, the proposed method depends primarily on the properties of the priors on the mixture probabilities and the mean vectors and not on the model for the covariance matrices. The mechanism can therefore be applied quite generally in conjunction with any model for the component specific covariance matrices belonging to a class that does not disturb the label invariance property of the posterior. This class includes both the MIW and the MLFA models described in Section 4.3.

4.4.2 Conditionally Heteroscedastic Measurement Errors

We now consider the case when the variances of the measurement errors depend on the associated unknown true values of the variables of interest through unknown relationships.

Consider the problem of flexible modeling of conditionally heteroscedastic regression errors where the response and the covariates are both univariate. Consider also the problem of modeling conditionally heteroscedastic measurement errors in a univariate deconvolution set up. As we have seen in Section 2, from a modeling perspective Bayesian hierarchical framework allows us to treat these two problems on par by treating both the covariate in the regression problem and the variable of interest in the deconvolution problem simply as conditioning variables. Of course in the regression problem X is precisely measured, whereas in the deconvolution problem X would be latent, but in either case we are required to flexibly model the density of $(U \mid X)$ subject to $E(U \mid X) = 0$, where U, depending upon the context, denotes either regression or measurement errors. Models for regression errors that allow their variance to vary with the values of the covariate (Pati and Dunson, 2013; Pelenis, 2014) can thus be tried as potential candidates for models for univariate conditionally heteroscedastic measurement errors. Conversely, the models for conditionally heteroscedastic univariate measurement errors (Staudenmayer, et al., 2008 and the ones we developed in Section 2) can also be employed to model univariate conditionally heteroscedastic regression errors.

This is not quite true in a multivariate set up. Interpreting the conditioning

variables $\mathbf{X}$ broadly as predictors, one can loosely connect the problem of modeling conditionally heteroscedastic multivariate measurement errors to the problem of covariance regression (Hoff and Niu, 2012; Fox and Dunson, 2013 etc.), where the covariance of multivariate regression errors are allowed to vary flexibly with precisely measured and possibly multivariate predictors. In covariance regression problems, the dimension of the regression errors is typically unrelated to the dimension of the predictors. In multivariate deconvolution problems, in contrast, the dimension of each $\mathbf{U}_{ij}$ is exactly the same as the dimension of $\mathbf{X}_i$, the k^{th} component U_{ijk} being the measurement error associated exclusively with X_{ik}. This implies that although different components of the measurement error vectors $\mathbf{U}_{ij}$ may be correlated, the dependence of U_{ijk} on $\mathbf{X}_i$ can be explained mostly through X_{ik}. For instance, the plots of subject specific means vs variances suggest that the conditional variability in U_{ijk} can be explained mostly as a function of X_{ik} only.

These characteristic differences between conditionally heteroscedastic multivariate regression and measurement errors preclude direct application of covariance regression approaches to model conditionally heteroscedastic measurement errors and warrant that models that can accommodate their unique characteristics be specially designed. As a plausible model for conditionally heteroscedastic multivariate measurement errors, we propose the following. We assume

$$(\mathbf{U}_{ij} \mid \mathbf{X}_i) = \mathbf{S}(\mathbf{X}_i)\boldsymbol{\epsilon}_{ij}, \tag{4.15}$$

where $\mathbf{S}(\mathbf{X}_i) = \text{diag}\{s_1(X_{i1}), s_2(X_{i2}), \ldots, s_p(X_{ip})\}$ and $\boldsymbol{\epsilon}_{ij}$, henceforth referred to as the 'scaled errors', are distributed independently of $\mathbf{X}_i$. Model (4.15) implies that $\text{cov}(\mathbf{U}_{ij} \mid \mathbf{X}_i) = \mathbf{S}(\mathbf{X}_i)\,\text{cov}(\boldsymbol{\epsilon}_{ij})\,\mathbf{S}(\mathbf{X}_i)$ and marginally $\text{var}(U_{ijk} \mid \mathbf{X}_i) = s_k^2(X_{ik})\text{var}(\epsilon_{ijk})$, a function of X_{ik} only, as desired. The techniques developed in Section 4.4.1 can now be employed to model the density of $\boldsymbol{\epsilon}_{ij}$, allowing different components of $\mathbf{U}_{ij}$ to be correlated and their joint density to deviate from multivariate normality.

We model the variance functions s_k^2, denoted also by v_k, using positive mixtures of B-spline basis functions with smoothness inducing priors on the coefficients as in Section 2. For the k^{th} component, we partition an interval $[A_k, B_k]$ of interest into L_k subintervals using knot points $A_k = t_{k,1} = \cdots = t_{k,q+1} < t_{k,q+2} < t_{k,q+3} < \cdots < t_{k,q+L_k} < t_{k,q+L_k+1} = \cdots = t_{k,2q+L_k+1} = B_k$. A flexible model for the variance

functions is given by

$$v_k(X_{ik}) = s_k^2(X_{ik}) = \sum_{j=1}^{J_k} b_{q,j,k}(X_{ik}) \exp(\xi_{jk}) = \mathbf{B}_{q,J_k,k}(X_{ik}) \exp(\boldsymbol{\xi}_k), \quad (4.16)$$

$$(\boldsymbol{\xi}_k \mid J_k, \sigma^2_{\xi,k}) \propto \exp\{-\boldsymbol{\xi}_k^{\mathrm{T}} P_k \boldsymbol{\xi}_k / (2\sigma^2_{\xi,k})\}, \quad \sigma^2_{\xi,k} \sim \text{Inv-Ga}(a_\xi, b_\xi). \quad (4.17)$$

Here $\{b_{q,j,k}\}_{j=1}^{J_k}$ denote $J_k = (q + L_k)$ B-spline bases of degree q as defined in de Boor (2000), $\boldsymbol{\xi}_k = \{\xi_{1k}, \xi_{2k}, \ldots, \xi_{J_k k}\}^{\mathrm{T}}$; $\exp(\boldsymbol{\xi}_k) = \{\exp(\xi_{1k}), \exp(\xi_{2k}), \ldots, \exp(\xi_{J_k k})\}^{\mathrm{T}}$; and $P_k = D_k^{\mathrm{T}} D_k$, where D_k is a $J_k \times (J_k + 2)$ matrix such that $D_k \boldsymbol{\xi}_k$ computes the second differences in $\boldsymbol{\xi}_k$.

4.5 Choice of Hyper-parameters

1. **Number of mixture components:** Practical application of our method requires that a decision be made on the number of mixture components $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$ in the models for the densities $f_{\mathbf{X}}$ and $f_{\boldsymbol{\epsilon}}$, respectively. Our simulation experiments suggest that when $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$ are assigned values greater than some minimum numbers required to approximate the target densities, the MCMC chain often quickly reaches a stable stage where the redundant components become empty. These observations are similar to that made in the context of ordinary density estimation by Rousseau and Mengersen (2011) who studied the asymptotic behavior of the posterior for overfitted mixture models and showed that when $\alpha/K < L/2$, where L denotes the number of parameters specifying the component kernels, the posterior is stable and concentrates in regions with empty redundant components. We set $\alpha_{\mathbf{X}} = \alpha_{\boldsymbol{\epsilon}} = 1$ so that the condition $\alpha/K < L/2$ is satisfied.

Educated guesses about $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$ may nevertheless be useful in safeguarding against gross overfitting that would result in a wastage of computation time and resources. The following simple strategies may be employed. Model based cluster analysis techniques as implemented by the mclust package in R may be applied to the starting values of $\mathbf{X}_i$ and the corresponding residuals, obtained by fitting univariate submodels for each component of $\mathbf{X}$, to get some idea about $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$. The chain may be started with larger values of $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$ and after a few hundred iterations the redundant empty components may be dumped on the fly.

As we will see in Section 4.8, the MIW method becomes highly numerically unstable when the measurement errors are conditionally heteroscedastic and the true

covariance matrices are highly sparse. In these cases in particular, the MIW method usually requires much larger sample sizes for the asymptotic results to hold and in finite samples the above mentioned strategy usually overestimates the required number of mixture components. Since mixtures based on $(K+1)$ components are at least as flexible as mixtures based on K components, as far as model flexibility is concerned, such overestimation is not an issue. But since this also results in clusters of smaller sizes, the estimates of the component specific covariance matrices become numerically even more unstable, further compounding the stability issues of the MIW model. In contrast, for the numerically more stable MLFA model, for the exact opposite reasons, the asymptotic results hold for much smaller sample sizes and such models are also more robust to overestimation of the number of nonempty clusters.

2. **Number of latent factors:** For the MLFA method, the MCMC algorithm summarized in Section 4.6 also requires that the component specific infinite factor models be truncated at some appropriate truncation level. The shrinkage prior again makes the model highly robust to overfitting allowing us to adopt a simple strategy. Since a latent factor characterization leads to a reduction in the number or parameters only when $q_k \leq \lfloor (p+1)/2 \rfloor$, we simply set the truncation level at $q_k = \lfloor (p+1)/2 \rfloor$ for all the components. We also experimented by setting the truncation level at $q_k = p$ for all k with the results remaining practically the same. The shrinkage prior being continuous in nature does not set the redundant columns to exact zeroes, but it adaptively shrinks the redundant parameters sufficiently towards zero producing stable and efficient estimates of the densities being modeled.

3. **Other hyper-parameters:** We take an empirical Bayes type approach to assign values to other hyper-parameters. We set $\boldsymbol{\mu}_{\mathbf{X},0} = \overline{\mathbf{X}}^{(0)}$, the overall mean of $\mathbf{X}^{(0)}_{1:n}$, where $\mathbf{X}^{(0)}_{1:n}$ denote the starting values of $\mathbf{X}_{1:n}$ for the MCMC sampler briefed in Section 4.6. For the scaled errors we set $\boldsymbol{\mu}_{\boldsymbol{\epsilon},0} = \mathbf{0}$. For the MIW model we take $\nu_0 = (p+2)$, the smallest possible integral value of ν_0 for which the prior mean of $\boldsymbol{\Sigma}_k$ exists. We then take $\boldsymbol{\Sigma}_{\mathbf{X},0}/2 = \boldsymbol{\Psi}_{\mathbf{X},0} = \text{cov}(\overline{\mathbf{X}}^{(0)}_{1:n})$. These choices imply $E(\boldsymbol{\Sigma}_{\mathbf{X},k}) = \boldsymbol{\Psi}_{\mathbf{X},0} = \text{cov}(\overline{\mathbf{X}}^{(0)})$ and, since the variability of each component is expected to be significantly less than the overall variability, ensure noninformativeness. Similarly, for the scaled errors we take $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon},0}/2 = \boldsymbol{\Psi}_{\boldsymbol{\epsilon},0} = \text{cov}(\boldsymbol{\epsilon}^{(0)}_{1:N})$. For the MLFA model, the hyper-parameters specifying the prior for $\boldsymbol{\Lambda}$ are set at $a_1 = 1, a_h = 2$ for all

$h \geq 2$, and $\nu = 1$. Inverse gamma priors with parameters $a_\sigma = 1.1, b_\sigma = 1$ are placed on the elements of $\boldsymbol{\Omega}$. For each k, the variance functions were modeled using quadratic (q=2) B-splines based on $(2 \times 2 + 5 + 1) = 10$ equidistant knot points on $[A_k, B_k] = [\min(\overline{\mathbf{W}}_{k,1:n}) - 0.1 \text{ range}(\overline{\mathbf{W}}_{k,1:n}), \max(\overline{\mathbf{W}}_{k,1:n}) + 0.1 \text{ range}(\overline{\mathbf{W}}_{k,1:n})]$, where $\overline{\mathbf{W}}_{k,1:n}$ denotes the subject specific means corresponding to k^{th} component.

4.6 Posterior Computation

Samples from the posterior can be drawn using a Gibbs sampler. In what follows $\boldsymbol{\zeta}$ denotes a generic variable that collects the observed proxies $\mathbf{W}_{1:N}$ and all the parameters of a model, including the imputed values of $\mathbf{X}_{1:n}$ and $\boldsymbol{\epsilon}_{1:N}$, that are not explicitly mentioned. To avoid unnecessary repetition, symbols sans the subscripts $\mathbf{X}$ and $\boldsymbol{\epsilon}$ are used as generics for similar components and parameters of the models. For example, $\boldsymbol{\mu}_k$ is a generic for $\boldsymbol{\mu}_{\mathbf{X},k}$ and $\boldsymbol{\mu}_{\boldsymbol{\epsilon},k}$; K is a generic for $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$; and so on.

Carefully chosen starting values can facilitate convergence of the sampler. The posterior means of the X_{ik}'s, obtained by fitting univariate versions of the proposed multivariate methods for each component of $\mathbf{X}$, are used as the starting values for the multivariate sampler. The number of mixture components are initialized at $K_{\mathbf{X}} = (m_{\mathbf{X}} + 2)$, where $m_{\mathbf{X}}$ denotes the optimal number of clusters returned by model based clustering algorithm implemented by the mclust package in R applied on the corresponding initial values $\mathbf{X}^{(0)}_{1:n}$. The component specific mean vectors of the nonempty clusters are set at the mean of $\mathbf{X}^{(0)}_i$ values that belong to that cluster. The component specific mean vectors of the two empty clusters are set at $\overline{\mathbf{X}}^{(0)}$, the overall mean of $\mathbf{X}^{(0)}_{1:n}$. For the MIW model, the initial values of the cluster specific covariance matrices are chosen in a similar fashion. The mixture probabilities for the k^{th} nonempty cluster is set at $\boldsymbol{\pi}_{\mathbf{X},k} = n_k/n$, where n_k denotes the number of $\mathbf{X}^{(0)}_i$ belonging to the k^{th} cluster. The mixture probabilities of the empty clusters are initialized at zero. For the MLFA method, the starting values of all elements of $\boldsymbol{\Lambda}$ and $\boldsymbol{\eta}$ are set at zero. The starting values for the elements of $\boldsymbol{\Omega}$ are chosen to equal the variances of the corresponding starting values. The parameters specifying the density of the scaled errors are initialized in a similar manner. The MCMC iterations comprise the following steps.

1. **Updating the parameters specifying the density of interest:** For the MIW model the parameters specifying the density $f_{\mathbf{X}}$ are updated using the following

steps.

$$\begin{aligned}
(\boldsymbol{\pi} \mid \boldsymbol{\zeta}) &\sim \text{Dir}(\alpha/K + n_1, \alpha/K + n_2, \ldots, \alpha/K + n_K) \\
(C_i \mid \boldsymbol{\zeta}) &\sim \text{Mult}(1, p_{i1}, p_{pi2}, \ldots, p_{iK}), \\
(\boldsymbol{\mu}_k \mid \boldsymbol{\zeta}) &\sim \text{MVN}_p(\boldsymbol{\mu}_k^{(n)}, \boldsymbol{\Sigma}_k^{(n)}), \\
(\boldsymbol{\Sigma}_k \mid \boldsymbol{\zeta}) &\sim \text{IW}_p\{n_k + \nu_0, \textstyle\sum_{i:C_i=k}(\mathbf{X}_i - \boldsymbol{\mu}_k)(\mathbf{X}_i - \boldsymbol{\mu}_k)^{\mathrm{T}} + \boldsymbol{\Psi}_0\},
\end{aligned}$$

where $n_k = \sum_i 1(C_i = k)$, $p_{ik} \propto \pi_k \times \text{MVN}_p(\mathbf{X}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $\boldsymbol{\Sigma}_k^{(n)} = (\boldsymbol{\Sigma}_0^{-1} + n_k\boldsymbol{\Sigma}_k^{-1})^{-1}$ and $\boldsymbol{\mu}_k^{(n)} = \boldsymbol{\Sigma}_k^{(n)}\left\{\boldsymbol{\Sigma}_k^{-1}\sum_{i:C_i=k}\mathbf{X}_i + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right\}$. To update the parameters specifying the covariance matrices in the MLFA model, the sampler cycles through the following steps.

$$\begin{aligned}
(\boldsymbol{\lambda}_{k,j} \mid \boldsymbol{\zeta}) &\sim \text{MVN}_q\{\boldsymbol{\Sigma}_{\boldsymbol{\lambda},jk}\sigma_j^{-2}\boldsymbol{\eta}_k^{\mathrm{T}}(\mathbf{X}_k^{(j)} - \boldsymbol{\mu}_k^{(j)}), \boldsymbol{\Sigma}_{\boldsymbol{\lambda},jk}\}, \\
(\boldsymbol{\eta}_i \mid C_i = k, \boldsymbol{\zeta}) &\sim \text{MVN}_q\{(\mathrm{I}_q + \boldsymbol{\Lambda}_k^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda}_k)^{-1}\boldsymbol{\Lambda}_k^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\mathbf{X}_i - \boldsymbol{\mu}_k), (\mathrm{I}_q + \boldsymbol{\Lambda}_k^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda}_k)^{-1}\}, \\
(\sigma_j^2 \mid \boldsymbol{\zeta}) &\sim \text{Inv-Ga}\left\{a_\sigma + n/2, b_\sigma + (1/2)\textstyle\sum_{i=1}^n(X_{ij} - \boldsymbol{\mu}_{C_i,j} - \boldsymbol{\lambda}_{C_i,j}^{\mathrm{T}}\boldsymbol{\eta}_i)^2\right\}, \\
(\phi_{k,jh} \mid \boldsymbol{\zeta}) &\sim \text{Ga}\{(\nu+1)/2, (\nu + \tau_{k,h}\lambda_{k,jh}^2)/2\}, \\
(\delta_{k,h} \mid \boldsymbol{\zeta}) &\sim \text{Ga}\{a_h + p(q-h+1)/2, 1 + \textstyle\sum_{\ell=1}^q \tau_{k,\ell}^{(h)}\sum_{j=1}^p \phi_{k,j\ell}\lambda_{k,j\ell}^2/2\},
\end{aligned}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\lambda},jk} = (\mathbf{D}_{k,j}^{-1} + \sigma_j^{-2}\boldsymbol{\eta}_k^{\mathrm{T}}\boldsymbol{\eta}_k)^{-1}$, $D_{k,j}^{-1} = \text{diag}(\phi_{k,j1}\tau_{k,1}, \ldots, \phi_{k,jq}\tau_{k,q})$, $\tau_{k,\ell}^{(h)} = \prod_{t=1,t\neq h}^{\ell}\delta_{k,t}$, $\mathbf{X}_k^{(j)} = (X_{i_1 j}, X_{i_2 j}, \ldots, X_{i_{n_k} j})^{\mathrm{T}}$, $\boldsymbol{\eta}_k^{n_k \times q} = (\boldsymbol{\eta}_{i_1}, \boldsymbol{\eta}_{i_2}, \ldots, \boldsymbol{\eta}_{i_{n_k}})^{\mathrm{T}}$, and $\{i : C_i = k\} = \{i_1, i_2, \ldots, i_{n_k}\}$.

2. **Updating the parameters specifying the density of the scaled errors:** Except one simple additional step that enforces the zero mean restriction, the steps to update the parameters specifying the error distribution are similar and thus excluded.

3. **Updating the values of the variables of interest:** When the measurement errors are independent of the variables of interest, $\mathbf{X}_i$ have closed form full conditionals given by

$$(\mathbf{X}_i \mid C_{\mathbf{X},i} = k, C_{\boldsymbol{\epsilon},i1} = k_1, \ldots, C_{\boldsymbol{\epsilon},im_i} = k_{m_i}, \boldsymbol{\zeta}) \sim \text{MVN}_p(\boldsymbol{\mu}_{\mathbf{X}}^{(n)}, \boldsymbol{\Sigma}_{\mathbf{X}}^{(n)}),$$

where $\boldsymbol{\Sigma}_{\mathbf{X}}^{(n)} = (\boldsymbol{\Sigma}_{\mathbf{X},k}^{-1} + \sum_{j=1}^{m_i}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon},k_j}^{-1})^{-1}$ and $\boldsymbol{\mu}_{\mathbf{X}}^{(n)} = \boldsymbol{\Sigma}_{\mathbf{X}}^{(n)}(\boldsymbol{\Sigma}_{\mathbf{X},k}^{-1}\boldsymbol{\mu}_{\mathbf{X},k} + \sum_{j=1}^{m_i}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon},k_j}^{-1}W_{ij})$.

For conditionally heteroscedastic errors, the full conditionals are given by

$$
\begin{aligned}
&(\mathbf{X}_i \mid C_{\mathbf{X},i} = k, C_{\boldsymbol{\epsilon},i1} = k_1, \ldots, C_{\boldsymbol{\epsilon},im_i} = k_{m_i}, \boldsymbol{\zeta}) \\
&\propto \mathrm{MVN}_p(\mathbf{X}_i \mid \boldsymbol{\mu}_{\mathbf{X},k}, \boldsymbol{\Sigma}_{\mathbf{X},k}) \times \textstyle\prod_{j=1}^{m_i} \mathrm{MVN}_p\{\mathbf{W}_{ij} \mid \mathbf{X}_i + \mathbf{S}(\mathbf{X}_i)\boldsymbol{\mu}_{\boldsymbol{\epsilon},k_j}, \mathbf{S}(\mathbf{X}_i)\boldsymbol{\Sigma}_{\boldsymbol{\epsilon},k_j}\mathbf{S}(\mathbf{X}_i)\},
\end{aligned}
$$

The full conditionals do not have closed forms. MH steps with multivariate normal random walk proposals are used within the Gibbs sampler.

4. **Updating the parameters specifying the variance functions:** When the measurement errors are conditionally heteroscedastic, we first estimate the variance functions $s_k^2(X_{ik})$ by fitting univariate submodels $W_{ijk} = X_{ik} + s_k(X_{ik})\epsilon_{ijk}$ for each k. The details are provided in Section 4.7. The parameters characterizing other components of the full model are then sampled using the Gibbs sampler described above, keeping the estimates of the variance functions fixed.

4.7 Estimation of the Variance Functions

When the measurement errors are conditionally heteroscedastic, we need to update the parameters $\boldsymbol{\xi}_k$ that specify the variance functions $s_k^2(X_{ik})$. These parameters do not have closed form full conditionals. MCMC algorithms, where we tried to integrate MH steps for $\boldsymbol{\xi}_k$ with the sampler for the parameters specifying $f_{\boldsymbol{\epsilon}}$, were numerically unstable. We need to supply the values of the scaled errors to step 2 of the algorithm described in Section 4.6 and the instability stems from the operation $\boldsymbol{\epsilon}_{ij} = \mathbf{S}(\mathbf{X}_i)^{-1}\mathbf{U}_{ij}$ required to calculate these values.

To solve the problem, we adopt a novel two-stage procedure. First, for each k, we estimate the functions $s_k^2(X_{ik})$ by fitting the univariate submodels $W_{ijk} = X_{ik} + s_k(X_{ik})\epsilon_{ijk}$. The problem of numerical instability arising out of the operation to determine the values of the scaled errors remains in these univariate subproblems too. But the following lemma from Pelenis (2014), presented here for easy reference, allows us to avoid this operation in the first place.

Lemma 1. *Let* $\boldsymbol{\theta}_{1:K} = \{(\pi_k, \mu_k, \sigma_k^2)\}_{k=1}^K$ *be such that*

$$
\begin{aligned}
&f_1(\epsilon \mid \boldsymbol{\theta}_{1:K}) = \textstyle\sum_{k=1}^K \pi_k \; \mathit{Normal}(\epsilon \mid \mu_k, \sigma_k^2), \\
&\mathit{with} \; \textstyle\sum_{k=1}^K \pi_k = 1, \; \mathit{and} \; \; \textstyle\sum_{k=1}^K \pi_k \mu_k = 0.
\end{aligned} \tag{4.18}
$$

Then there exists a set of parameters $\boldsymbol{\theta}^{\star}_{1:(K-1)}$ *such that*

$$f_1(\epsilon \mid \boldsymbol{\theta}_{1:K}) = f_2(\epsilon \mid \boldsymbol{\theta}^{\star}_{1:(K-1)}) = \sum_{k=1}^{K-1} \pi^{\star}_k \sum_{\ell=1}^{2} p^{\star}_{k,\ell} Normal(\epsilon \mid \mu^{\star}_{k,\ell}, \sigma^{\star 2}_{k,\ell}), \quad (4.19)$$
$$\sum_{k=1}^{K-1} \pi^{\star}_k = 1, \quad \sum_{\ell=1}^{2} p^{\star}_{k,\ell} = 1, \quad \sum_{\ell=1}^{2} p^{\star}_{k,\ell} \mu^{\star}_{k,\ell} = 0 \ \forall k.$$

Lemma 1 implies that the univariate submodels for the density of the scaled errors given by (4.18) has a reparametrization (4.19) where each component is itself a two-component normal mixture with its mean restricted at zero. The reparametrization (4.19) thus replaces the zero mean restriction on (4.18) by similar restrictions on each of its components. These restrictions also imply that each mixture component in (4.19) can be characterized by only four free parameters. The model for the scaled errors used in the univariate density deconvolution problem discussed in Section 2 of this dissertation was essentially an infinite dimensional extension of this reparametrized finite dimensional submodel where we allowed $K_\epsilon \to \infty$ and $\boldsymbol{\pi}_\epsilon \sim \text{Stick}(\alpha_\epsilon)$. The problem of numerical instability was tackled by using MH steps to update not only the parameters specifying the variance function but also the parameters characterizing the density of the scaled errors using the conditional likelihood $f_{U|X}$ (and not f_ϵ), thus escaping the need to separately determine the values of the scaled errors. See the details given in Section 2.5. The same can also be done for the finite dimensional submodels here. High precision estimates of the variance functions can be obtained using these reparametrized finite dimensional univariate deconvolution models. See Figure 4.1 in this section and Figure C.1 in Appendix C for illustrations. The priors and the hyper-parameters for the univariate submodels can be chosen following the suggestions given in Secion 2. The strategy of exploiting the properties of overfitted mixture models to determine the number of mixture components described in Section 4.5 can also be applied to the univariate subproblems.

A similar reparametrization does exist for the multivariate problem too, but the strategy would not be very effective in the multivariate set up as it would require updating the mean vectors and the covariance matrices involved in $f_{\boldsymbol{\epsilon}}$ through MH steps which are not efficient in simultaneous updating of large numbers of parameters. After estimating the functions variance functions from the univariate submodels, we therefore keep these estimates fixed and sample other parameters using the Gibbs sampler described in Section 4.6.

4.8 Simulation Experiments

We recall that the mean integrated squared error (MISE) of estimation of $f_{\mathbf{X}}$ by $\widehat{f}_{\mathbf{X}}$ is defined as $MISE = E\int\{f_{\mathbf{X}}(\mathbf{X}) - \widehat{f}_{\mathbf{X}}(X)\}^2 dX$. In the multivariate problem it becomes computationally extremely intensive to compute the MISE by estimating $f_{\mathbf{X}}$ on high dimensional equi-spaced grids. We therefore use importance sampling techniques. Based on B simulated data sets, a Monte Carlo estimate of MISE is given by $MISE_{est} = B^{-1}\sum_{b=1}^{B}\sum_{m=1}^{M}\{f_{\mathbf{X}}(\mathbf{X}_{b,m}) - \widehat{f}_{\mathbf{X}}^{(b)}(\mathbf{X}_{b,m})\}^2/p_0(\mathbf{X}_{b,m})$, where $\{\mathbf{X}_{b,m}\}_{b=1,m=1}^{B,M}$ are random samples from the density p_0. We designed simulation experiments to evaluate the MISE performance of the proposed models for a wide range of possibilities. The MISEs we report here are all based on 100 simulated data sets and $M = 10^6$ samples generated from each of the two densities (a) $p_0 = f_{\mathbf{X}}$, the true density of $\mathbf{X}$, and (b) p_0 that is uniform on the hypercube with edges $\min_k\{\boldsymbol{\mu}_{\mathbf{X},k} - 3\mathbf{1}_p\}$ and $\max_k\{\boldsymbol{\mu}_{\mathbf{X},k} + 3\mathbf{1}_p\}$. The choice of prior hyper-parameters and details of the MCMC algorithm used to draw samples from the posterior are presented in the Appendix. With carefully chosen initial values and proposal densities for the MH steps, we were able to achieve quick convergence for the MCMC samplers. The use of symmetric Dirichlet priors helped simply mixing issues (Geweke, 2007). See Appendix C.1.1 for additional discussions. We programmed in R. In each case, we ran 3000 MCMC iterations and discarded the initial 1000 iterations as burn-in. For the univariate samplers, 1000 MCMC iterations with a burn-in of 500 sufficed to produce stable estimates of the variance functions. The post burn-in samples were thinned by a thinning interval of length 5. In our experiments with much larger iteration numbers and burn-ins, the MISE performances remained practically the same. This being the first article that tries to solve the problem of multivariate density deconvolution when the measurement error density is unknown, the proposed MIW and MLFA models had essentially no competitors. We thus compared our models with a naive Bayesian method that ignores measurement errors and treats the subject specific means as precisely measured observations instead, modeling the density of interest by a finite mixture of multivariate normals as in (4.2) with inverse Wishart priors on the component specific covariance matrices.

We considered two choices for the sample size $n = 500, 1000$. For each subject, we simulated $m_i = 3$ replicates. The true density of interest $f_{\mathbf{X}}$ was chosen to be $f_{\mathbf{X}}(\mathbf{X}) = \sum_{k=1}^{K_{\mathbf{X}}}\pi_{\mathbf{X},k}\ \mathrm{MVN}_p(\mathbf{X} \mid \boldsymbol{\mu}_{\mathbf{X},k}, \boldsymbol{\Sigma}_{\mathbf{X},k})$ with $K_{\mathbf{X}} = 3$, $\boldsymbol{\pi}_{\mathbf{X}} = (0.25, 0.50, 0.25)^{\mathrm{T}}$, $\boldsymbol{\mu}_{\mathbf{X},1} = (0.8, 6, 4, 5)^{\mathrm{T}}$, $\boldsymbol{\mu}_{\mathbf{X},2} = (2.5, 4, 5, 6)^{\mathrm{T}}$ and $\boldsymbol{\mu}_{\mathbf{X},3} = (6, 4, 2, 4)^{\mathrm{T}}$. For the density

of the measurement errors $f_{\boldsymbol{\epsilon}}$ we considered two choices, namely

1. $f^1_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) = \text{MVN}_p(\boldsymbol{\epsilon} \mid \mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, and

2. $f^2_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) = \sum_{k=1}^{K_{\boldsymbol{\epsilon}}} \pi_{\boldsymbol{\epsilon},k}\ \text{MVN}_p(\boldsymbol{\epsilon} \mid \boldsymbol{\mu}_{\boldsymbol{\epsilon},k}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon},k})$ with $K_{\boldsymbol{\epsilon}} = 3$, $\boldsymbol{\pi}_{\boldsymbol{\epsilon}} = (0.2, 0.6, 0.2)^{\text{T}}$, $\boldsymbol{\mu}_{\boldsymbol{\epsilon},1} = (-0.3, 0, 0.3, 0)^{\text{T}}$, $\boldsymbol{\mu}_{\boldsymbol{\epsilon},2} = (-0.5, 0.4, 0.5, 0)^{\text{T}}$ and $\boldsymbol{\mu}_{\boldsymbol{\epsilon},3} = -(\pi_{\boldsymbol{\epsilon},1}\boldsymbol{\mu}_{\boldsymbol{\epsilon},1} + \pi_{\boldsymbol{\epsilon},2}\boldsymbol{\mu}_{\boldsymbol{\epsilon},2})/\pi_{\boldsymbol{\epsilon},3}$.

For the component specific covariance matrices, we set $\boldsymbol{\Sigma}_{\mathbf{X},k} = \mathbf{D}_{\mathbf{X}}\boldsymbol{\Sigma}_{\mathbf{X},0}\mathbf{D}_{\mathbf{X}}$ for each k, where $\mathbf{D}_{\mathbf{X}} = \text{diag}(0.75^{1/2}, \ldots, 0.75^{1/2})$. Similarly, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon},k} = \mathbf{D}_{\boldsymbol{\epsilon}}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon},0}\mathbf{D}_{\boldsymbol{\epsilon}}$ for each k, where $\mathbf{D}_{\boldsymbol{\epsilon}} = \text{diag}(0.3^{1/2}, \ldots, 0.3^{1/2})$. For each pair of $f_{\mathbf{X}}$ and $f_{\boldsymbol{\epsilon}}$, we considered four types of covariance structures for $\boldsymbol{\Sigma}_{\mathbf{X},0} = ((\sigma_{ij}^{\mathbf{X},0}))$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon},0} = ((\sigma_{ij}^{\boldsymbol{\epsilon},0}))$, namely

1. Identity (I): $\boldsymbol{\Sigma}_{\mathbf{X},0} = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon},0} = \text{I}_p$,

2. Latent Factor (LF): $\boldsymbol{\Sigma}_{\mathbf{X},0} = \boldsymbol{\Lambda}_{\mathbf{X}}\boldsymbol{\Lambda}_{\mathbf{X}} + \boldsymbol{\Omega}_{\mathbf{X}}$, with $\boldsymbol{\Lambda}_{\mathbf{X}} = (0.7, \ldots, 0.7)^{\text{T}}$ and $\boldsymbol{\Omega}_{\mathbf{X}} = \text{diag}(0.51, \ldots, 0.51)$, and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon},0} = \boldsymbol{\Lambda}_{\boldsymbol{\epsilon}}\boldsymbol{\Lambda}_{\boldsymbol{\epsilon}} + \boldsymbol{\Omega}_{\boldsymbol{\epsilon}}$, with $\boldsymbol{\Lambda}_{\boldsymbol{\epsilon}} = (0.5, \ldots, 0.5)^{\text{T}}$ and $\boldsymbol{\Omega}_{\boldsymbol{\epsilon}} = \text{diag}(0.75, \ldots, 0.75)$,

3. Autoregressive (AR): $\sigma_{ij}^{\mathbf{X},0} = 0.7^{|i-j|}$ and $\sigma_{ij}^{\boldsymbol{\epsilon},0} = 0.5^{|i-j|}$ for each (i, j), and

4. Exponential (EXP): $\sigma_{ij}^{\mathbf{X},0} = \exp(-0.5\,|i-j|)$ and $\sigma_{ij}^{\boldsymbol{\epsilon},0} = \exp(-0.9\,|i-j|)$ for each (i, j).

The parameters were chosen to produce a wide variety of one and two dimensional marginal densities. Scale adjustments by multiplication with $\mathbf{D}_{\mathbf{X}}$ and $\mathbf{D}_{\boldsymbol{\epsilon}}$ were done so that the simulated values of each component of $\mathbf{X}$ fall essentially in the range $(-2, 6)$ and the simulated values of all components of $\boldsymbol{\epsilon}$ fall essentially in the range $(-3, 3)$. For conditionally heteroscedastic measurement errors, we set the true variance functions at $s_k^2(X) = (1+X/4)^2$ for each component k. A total of 32 ($2\times1\times4\times4$) cases were thus considered for both independent and conditionally heteroscedastic measurement errors.

We first discuss the results of the simulation experiments when the measurement errors $\mathbf{U}$ were independent of the vector of interest $\mathbf{X}$. The estimated MISEs are presented in Table 4.1. When the true measurement error density was a single component multivariate normal, the MLFA model produced the lowest MISE when the true covariance matrices were diagonal. In all other cases the MIW model produced

the best results. When the true measurement error density was a mixture of multivariate normals, the model complexity increases and the performance of the MIW model started to deteriorate. In this case, the MLFA model dominated the MIW model when the true covariance matrices were either diagonal or had a latent factor characterization.

The estimated MISEs for the cases when the measurement errors were conditionally heteroscedastic are presented in Table 4.2. Models that accommodate conditional heteroscedasticity are significantly more complex compared to models that assume independence of the measurement errors from the vector of interest. The numerically more stable MLFA model thus out-performed the MIW model in all 32 cases. The improvements were particularly significant when the true covariance matrices were sparse and the number of subjects was small ($n = 500$). The true and estimated univariate and bivariate marginals for the density of interest $f_{\mathbf{X}}$ produced by the MIW and the MLFA methods when the true density of the scaled errors was a mixture of multivariate normals ($f^2_{\boldsymbol{\epsilon}}$) and the component specific covariance matrices were diagonal (I) are summarized in Figure 4.2 and Figure 4.3, respectively. The true and estimated univariate and bivariate marginals for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$ for this case produced by the two methods are summarized in Figure 4.4 and Figure 4.5, respectively. The true and the estimated variance functions produced by the univariate submodels are summarized in Figure 4.1. Comparisons between Figure 4.2 and Figure 4.3 illustrate the limitations of the MIW models in capturing high dimensional sparse covariance matrices and the improvements that can be achieved by the MLFA models. The estimates of $f_{\boldsymbol{\epsilon}}$ produced by the two methods agree more. This may be attributed to the fact that many more residuals are available for estimating $f_{\boldsymbol{\epsilon}}$ than there are $\mathbf{X}_i$'s to estimate $f_{\mathbf{X}}$. Figure 4.1 shows that the univariate submodels can recover the true variance functions well. Figure 4.6 and Figure 4.7 show the trace plots of the number of non-empty mixture components for diagonal covariance matrices for the MIW and the MLFA model, respectively. As discussed in Section 4.5, for highly sparse covariance matrices, the MIW model over-estimates the number of non-empty mixture components. The MLFA model, on the other hand, is much more robust to misspecification of the number of mixture components and correctly estimates the true number of mixture models. Additional figures when the true covariance matrices had auto-regressive structure (AR) are presented in Appendix C. In this case the true covariance matrices were not sparse.

The MLFA method still vastly dominated the MIW method when the sample size was small ($n = 500$). When the sample size was large ($n = 1000$) the two methods produced comparable results.

True Error Distribution	Covariance Structure	Sample Size	MISE $\times 10^4$		
			MLFA	MIW	Naive
(a) Multivariate Normal	I	500	**1.24**	3.05	8.01
		1000	**0.59**	1.33	6.58
	LF	500	6.88	**6.33**	33.41
		1000	5.15	**3.10**	32.42
	AR	500	11.91	**5.51**	27.17
		1000	9.82	**2.78**	26.01
	EXP	500	7.15	**4.40**	17.82
		1000	5.46	**2.19**	17.40
(b) Mixture of Multivariate Normal	I	500	**1.28**	3.24	5.97
		1000	**0.64**	1.37	4.99
	LF	500	**7.28**	7.51	31.62
		1000	**4.17**	4.34	31.48
	AR	500	10.43	**6.66**	30.74
		1000	7.75	**4.35**	28.90
	EXP	500	7.16	**5.18**	17.85
		1000	4.87	**2.66**	17.26

Table 4.1: Mean integrated squared error (MISE) performance of MLFA (mixtures of latent factor analyzers) and MIW (mixtures with inverse Wishart priors) density deconvolution models described in Section 4.2 of this dissertation for homoscedastic errors compared with a naive method that ignores measurement errors for different measurement error distributions. The minimum value in each row is highlighted.

4.9 Example

The Eating at America's Table study (EATS) (Subar, et al., 2001) conducted by the National Cancer Institute (NCI), that we briefly described in Section 2, again serves as the motivation for the methodology developed in this section. We recall that in this study $n = 965$ participants were interviewed $m_i = 4$ times over the course of a year and their 24 hour dietary recalls ($\mathbf{W}_{ij}$'s) were recorded. In section 2, we considered the problem of estimating the consumption patterns of individual dietary components. But nutritionists are typically interested not just in the consumption patterns of individual dietary components but also in their joint consumption pat-

True Error Distribution	Covariance Structure	Sample Size	MISE $\times 10^4$		
			MLFA	MIW	Naive
(a) Multivariate Normal	I	500	**2.53**	19.08	10.64
		1000	**1.15**	9.43	9.14
	LF	500	**11.46**	34.21	21.33
		1000	**5.78**	15.98	20.75
	AR	500	**17.11**	30.83	36.44
		1000	**10.77**	12.46	36.37
	EXP	500	**11.63**	26.99	24.28
		1000	**6.67**	10.56	23.36
(b) Mixture of Multivariate Normal	I	500	**2.79**	22.17	20.16
		1000	**1.38**	10.55	19.39
	LF	500	**13.39**	35.67	43.43
		1000	**7.50**	20.86	43.28
	AR	500	**18.27**	35.70	75.26
		1000	**12.06**	16.64	77.55
	EXP	500	**12.11**	34.50	48.76
		1000	**7.59**	13.74	50.02

Table 4.2: Mean integrated squared error (MISE) performance of MLFA (mixtures of latent factor analyzers) and MIW (mixtures with inverse Wishart priors) density deconvolution models described in Section 4.2 of this dissertation for conditionally heteroscedastic errors compared with a naive method that ignores measurement errors or different measurement error distributions. The minimum value in each row is highlighted.

terns. The goal of this section is to estimate the joint consumption patterns of the true daily intakes ($\mathbf{X}_i$'s) from their contaminated 24-hour recalls.

To illustrate our methodology, we consider the problem of estimating the joint consumption pattern of four dietary components, namely carbohydrate (1), fiber (2), protein (3) and a mineral potassium (4). Figure 4.8 shows the estimates of the variance functions produced by univariate submodels superimposed over plots of subject-specific means versus subject-specific variances for daily intakes of the dietary components. As is clearly identifiable from this plot, conditional heteroscedasticity is a very prominent feature of the measurements errors contaminating the 24 hour recalls. The estimated univariate and bivariate marginal densities of average long term daily intakes of the dietary components produced by the MIW method and the MLFA method are summarized in Figure 4.9. The estimated univariate and bivariate marginal densities for the scaled errors are summarized in Figure 4.10. The estimated marginals for the variables of interest produced by the two methods look

quite different, while the estimated marginals of the density of the scaled errors are in close agreement. The estimated univariate and bivariate marginal densities of the long term intakes of the dietary components produced by the MIW model look very irregular and unstable, whereas the estimates produced by the MLFA model look relatively more regular and stable. In experiments, not reported here, we observed that the estimates produced by the MIW method were sensitive to the choice of the number of mixture components, but the estimates produced by the MLFA model were quite robust. The trace plots and the frequency distributions of the of the numbers of nonempty mixture components are summarized in Figure 4.11 and Figure 4.12 and provide some idea about the relative stability of the two methods. These observations are similar to that made in Section 4.8 for conditionally heteroscedastic measurement errors and sparse covariance matrices.

4.10 Conclusion

In this section we considered the problem of multivariate density deconvolution when the measurement error density is not known but replicated proxies are available for unknown value of the vector valued variable of interest. We proposed Bayesian semiparametric solutions for two types of scenarios: 1. when the measurement errors are distributed independently of the variable of interest, and 2. when the variability of different components of the measurement errors depends on the associated unknown value of the variable of interest through unknown relationship. We used flexible finite mixtures of multivariate normal kernels with symmetric Dirichlet priors on the mixture probabilities to model both the density of interest and the density of the measurement errors. Utilizing the symmetry of the prior on the mixture probabilities and basic properties of multivariate normal distributions and finite mixture models, we proposed a novel technique to make the mixture model for the density of the errors satisfy a zero mean restriction. We showed that the very presence of measurement errors can compound numerical stability issues and make the dense parametrization of inverse Wishart priors unsuitable for modeling the component specific covariance matrices of the mixture models even in small dimensions, particularly in the case of conditionally heteroscedastic measurement errors. We proposed an alternative approach based on latent factor characterization of the covariance matrices with sparsity inducing priors on the factor loading matrices that led to significantly better performance in those scenarios. We built models for conditionally heteroscedastic

measurement errors taking into account their somewhat unique characteristics and proposed a novel two stage procedure to tackle the computational obstacles associated with such complicated high dimensional models using reparametrized versions of the corresponding univariate submodels. We illustrated the efficiency of the proposed methods in recovering the true density of the vector valued variable of interest through simulation experiments. Our work was motivated mostly by problems in nutritional epidemiology, but the methods we described addressed a broad topic of statistical research and should find potential applications in many other fields of applied research. To the best of our knowledge, all existing multivariate deconvolution methods assume the measurement error density to be fully specified. By allowing the density of the measurement errors to be unknown and free from parametric laws and also accommodating conditional heteroscedasticity, the methodology developed in section thus makes important contributions to the sparse literature on the problem of multivariate density deconvolution.

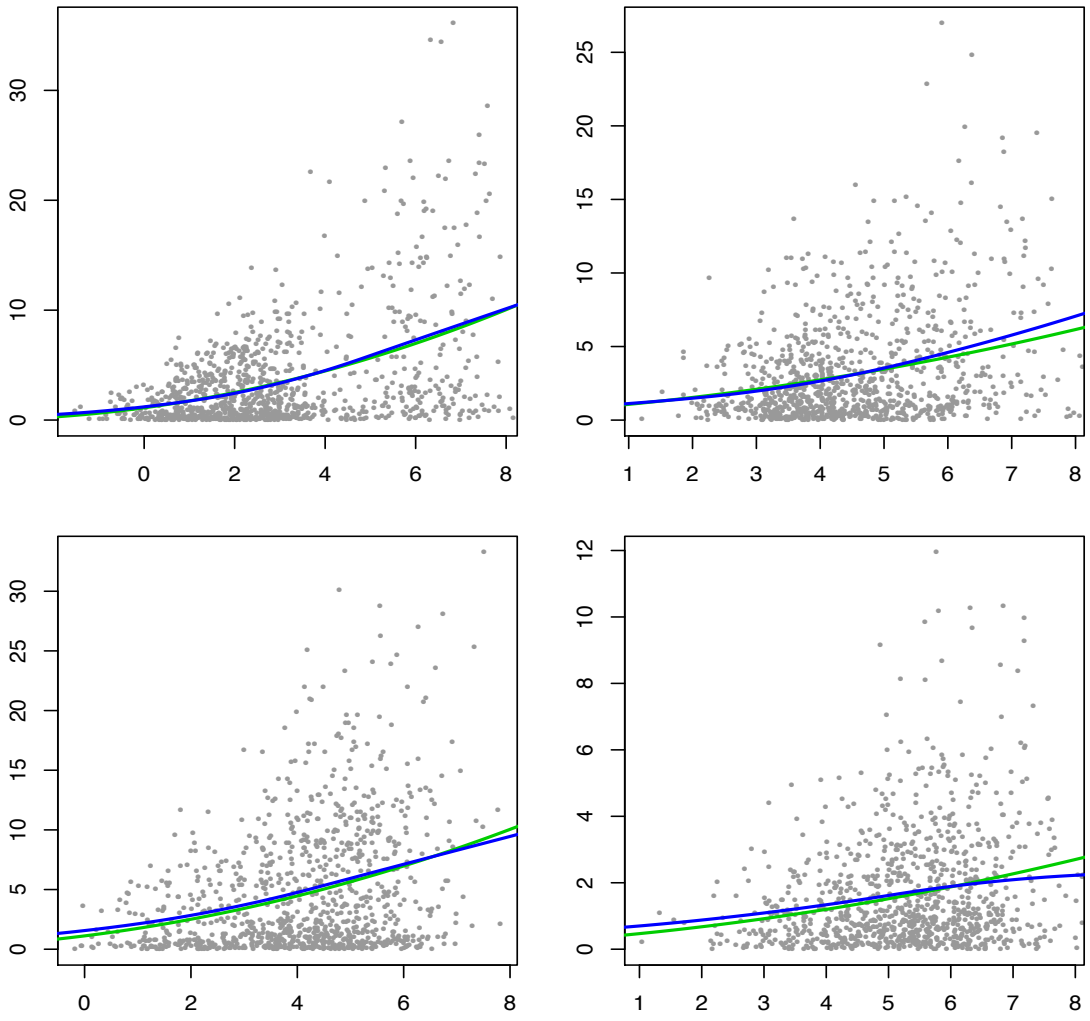

Figure 4.1: Results for the variance functions $s^2(X)$ produced by the univariate density deconvolution method for each component of the vector of interest $\mathbf{X}$ for the conditionally heteroscedastic error distribution $f^2_{\boldsymbol{\epsilon}}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and identity matrix (I) for the component specific covariance matrices. The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets for the MLFA (mixtures of latent factor analyzers) method. For each component of $\mathbf{X}$, the true variance function is $s^2(X) = (1 + X/4)^2$. See Section 4.4.2 and 4.7 for additional details. In each panel, the true (lighter shaded lines) and the estimated (darker shaded lines) variance functions are superimposed over a plot of subject specific sample means vs subject specific sample variances. The figure is in color in the electronic version of this dissertation.

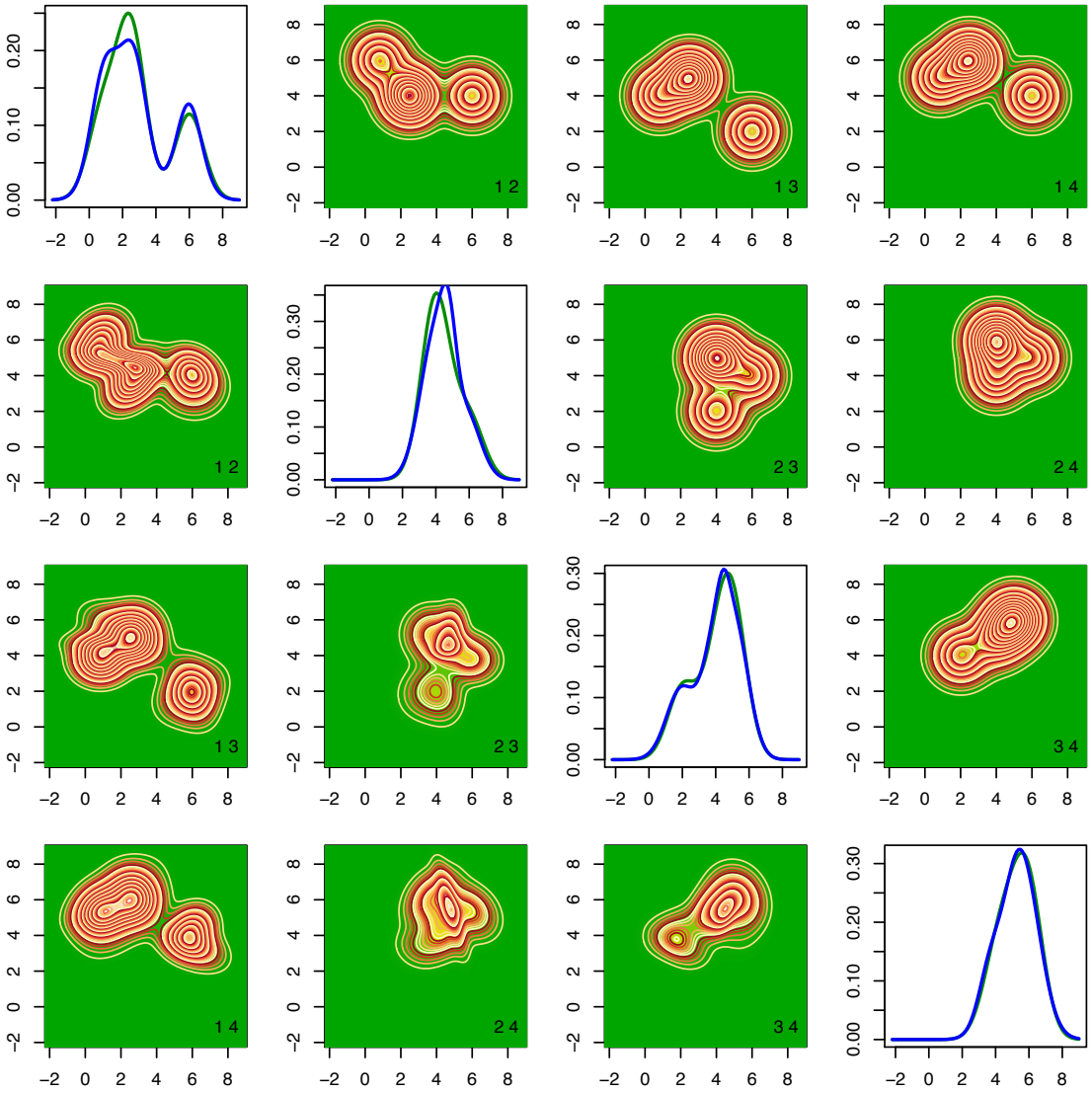

Figure 4.2: Results for the density of interest $f_{\mathbf{X}}$ produced by the MIW (mixtures with inverse Wishart priors) method for the conditionally heteroscedastic error distribution $f^2_{\boldsymbol{\epsilon}}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and identity matrix (I) for the component specific covariance matrices. The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets. See Section 4.8 for additional details. The upper triangular panels show the contour plots of the true two dimensional marginal densities. The lower triangular diagonally opposite panels show the corresponding estimates. The diagonal panels show the true (lighter shaded lines) and the estimated (darker shaded lines) one dimensional marginals. The figure is in color in the electronic version of this dissertation.

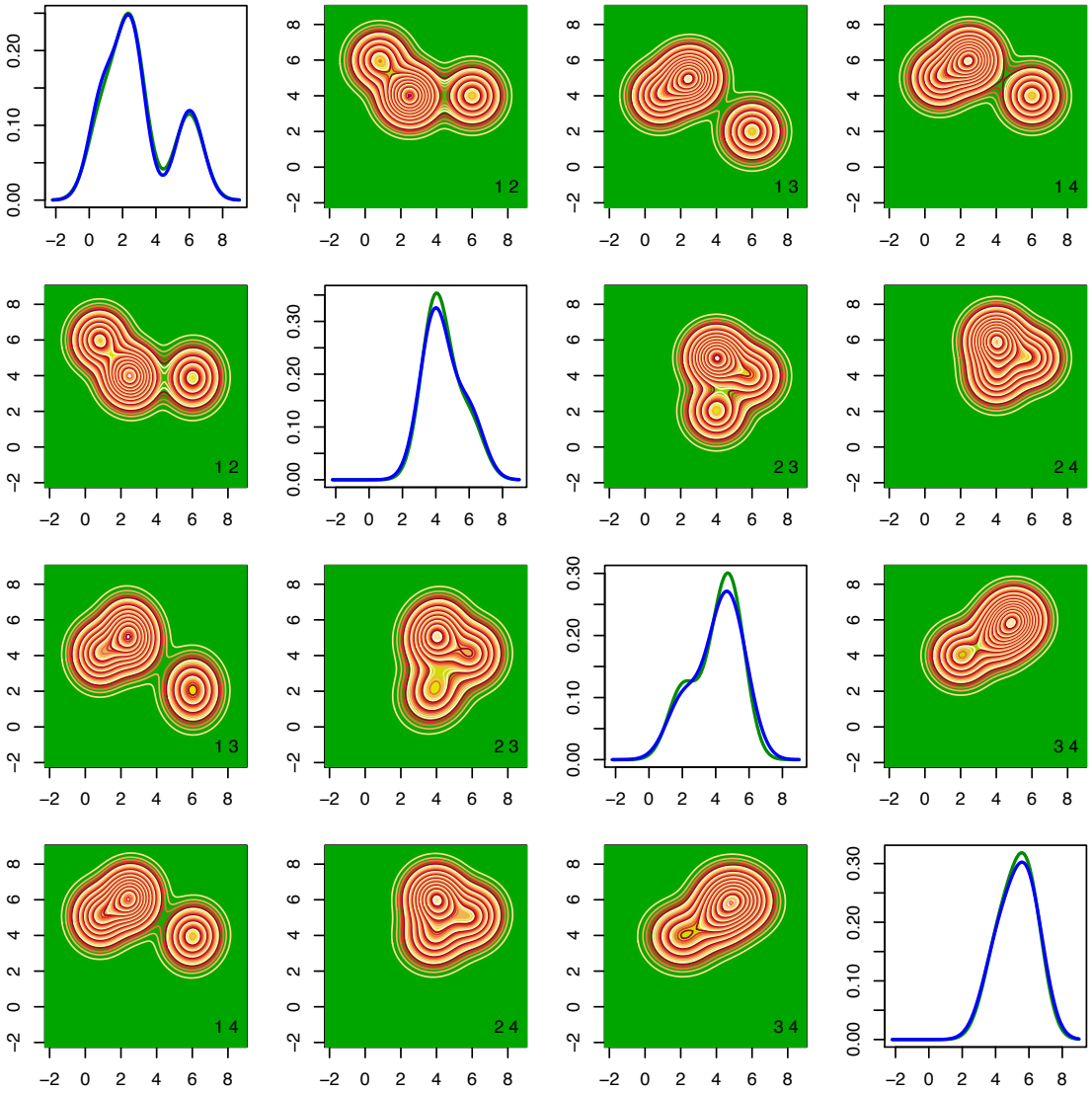

Figure 4.3: Results for the density of interest $f_{\mathbf{X}}$ produced by the MLFA (mixtures of latent factor analyzers) method for the conditionally heteroscedastic error distribution $f^2_{\boldsymbol{\epsilon}}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and identity matrix (I) for the component specific covariance matrices. The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets. See Section 4.8 for additional details. The upper triangular panels show the contour plots of the true two dimensional marginal densities. The lower triangular diagonally opposite panels show the corresponding estimates. The diagonal panels show the true (lighter shaded lines) and the estimated (darker shaded lines) one dimensional marginals. The figure is in color in the electronic version of this dissertation.

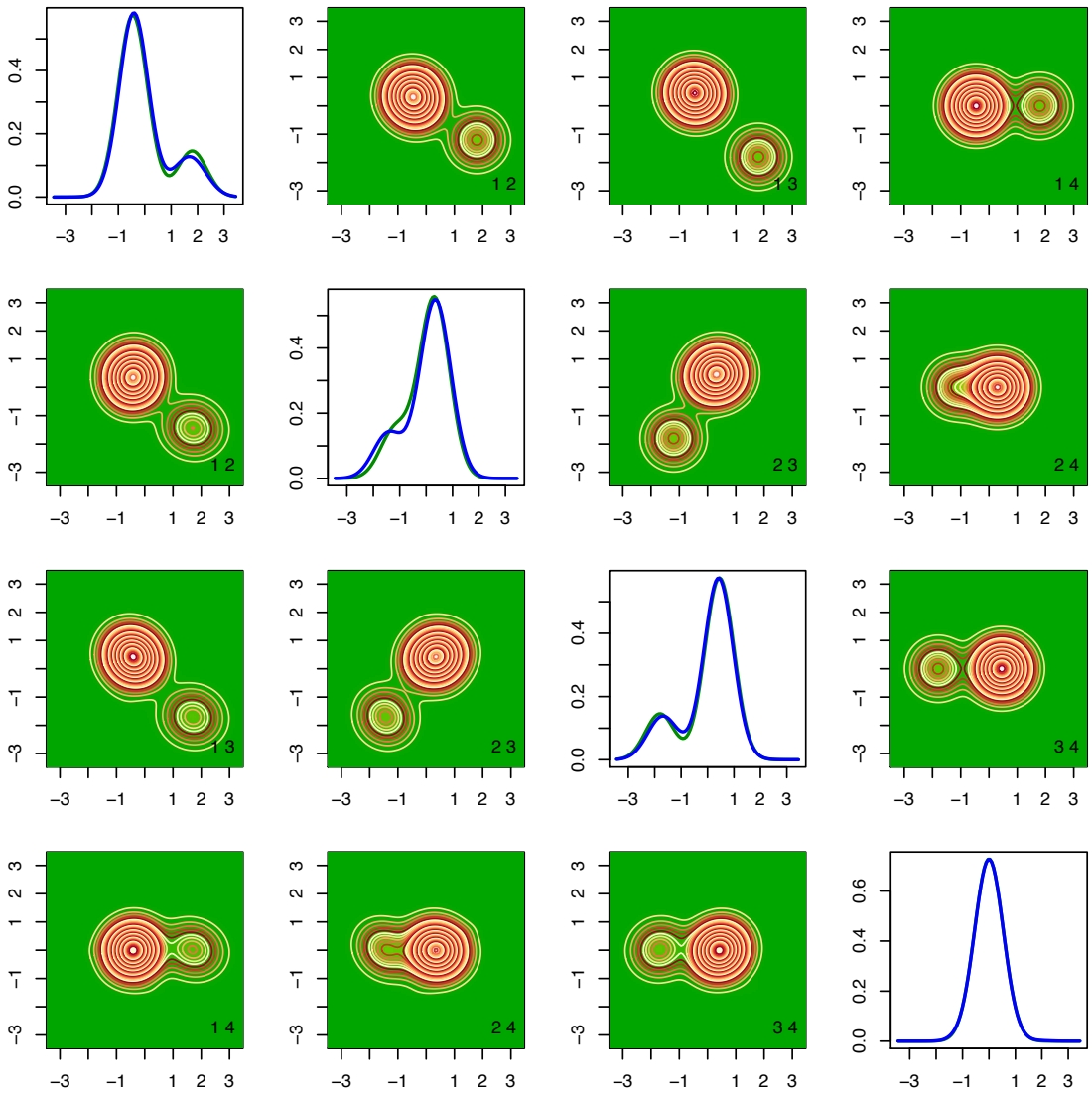

Figure 4.4: Results for the density of the scaled measurement errors $f_{\boldsymbol{\epsilon}}$ produced by the MIW (mixtures with inverse Wishart priors) method for the conditionally heteroscedastic error distribution $f^2_{\boldsymbol{\epsilon}}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and identity matrix (I) for the component specific covariance matrices. The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets. See Section 4.8 for additional details. The upper triangular panels show the contour plots of the true two dimensional marginal densities. The lower triangular diagonally opposite panels show the corresponding estimates. The diagonal panels show the true (lighter shaded lines) and the estimated (darker shaded lines) one dimensional marginals. The figure is in color in the electronic version of this dissertation.

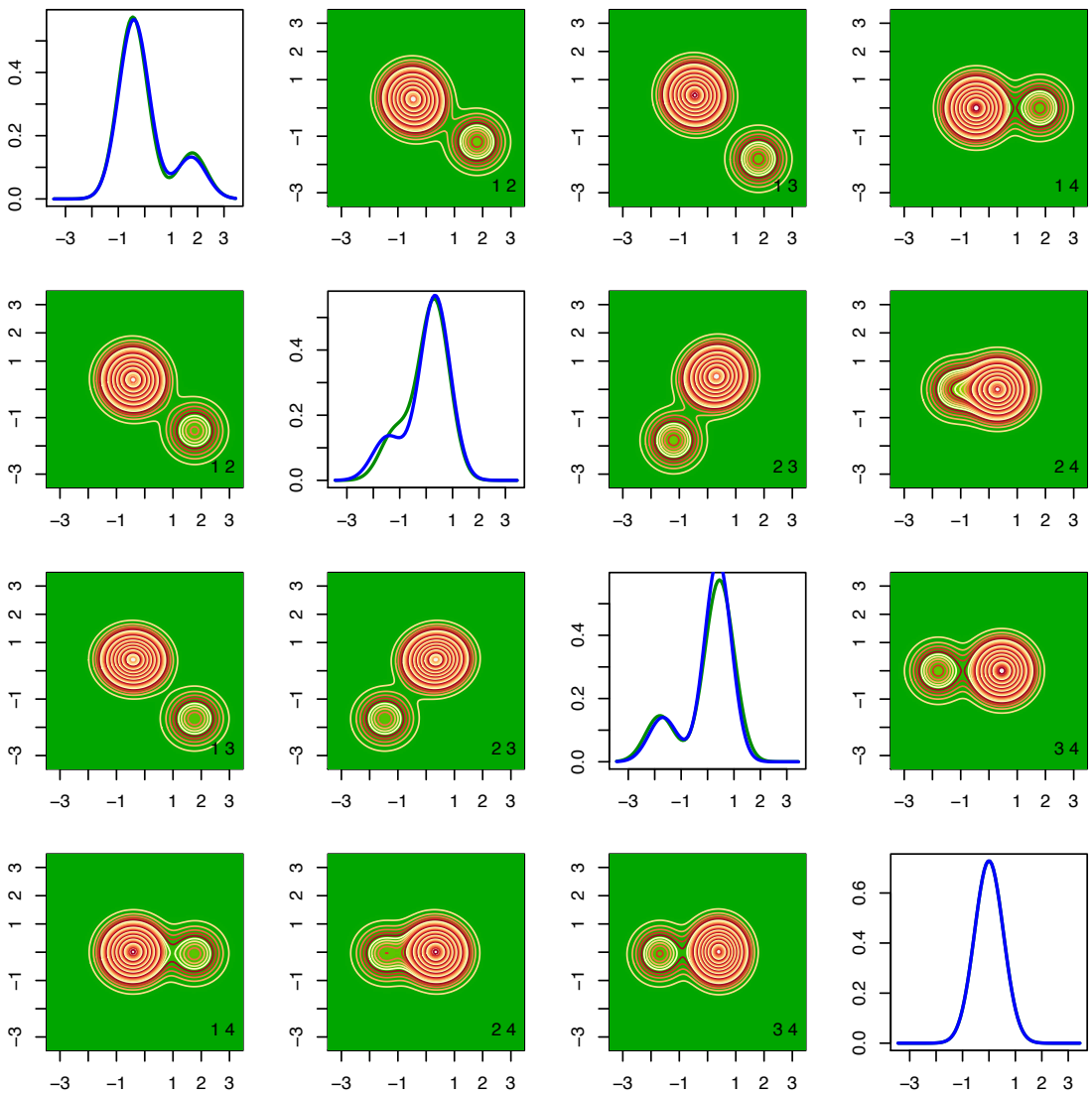

Figure 4.5: Results for the density of the scaled measurement errors $f_{\boldsymbol{\epsilon}}$ produced by the MLFA (mixtures of latent factor analyzers) method for the conditionally heteroscedastic error distribution $f^2_{\boldsymbol{\epsilon}}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and identity matrix (I) for the component specific covariance matrices. The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets. See Section 4.8 for additional details. The upper triangular panels show the contour plots of the true two dimensional marginal densities. The lower triangular diagonally opposite panels show the corresponding estimates. The diagonal panels show the true (lighter shaded lines) and the estimated (darker shaded lines) one dimensional marginals. The figure is in color in the electronic version of this dissertation.

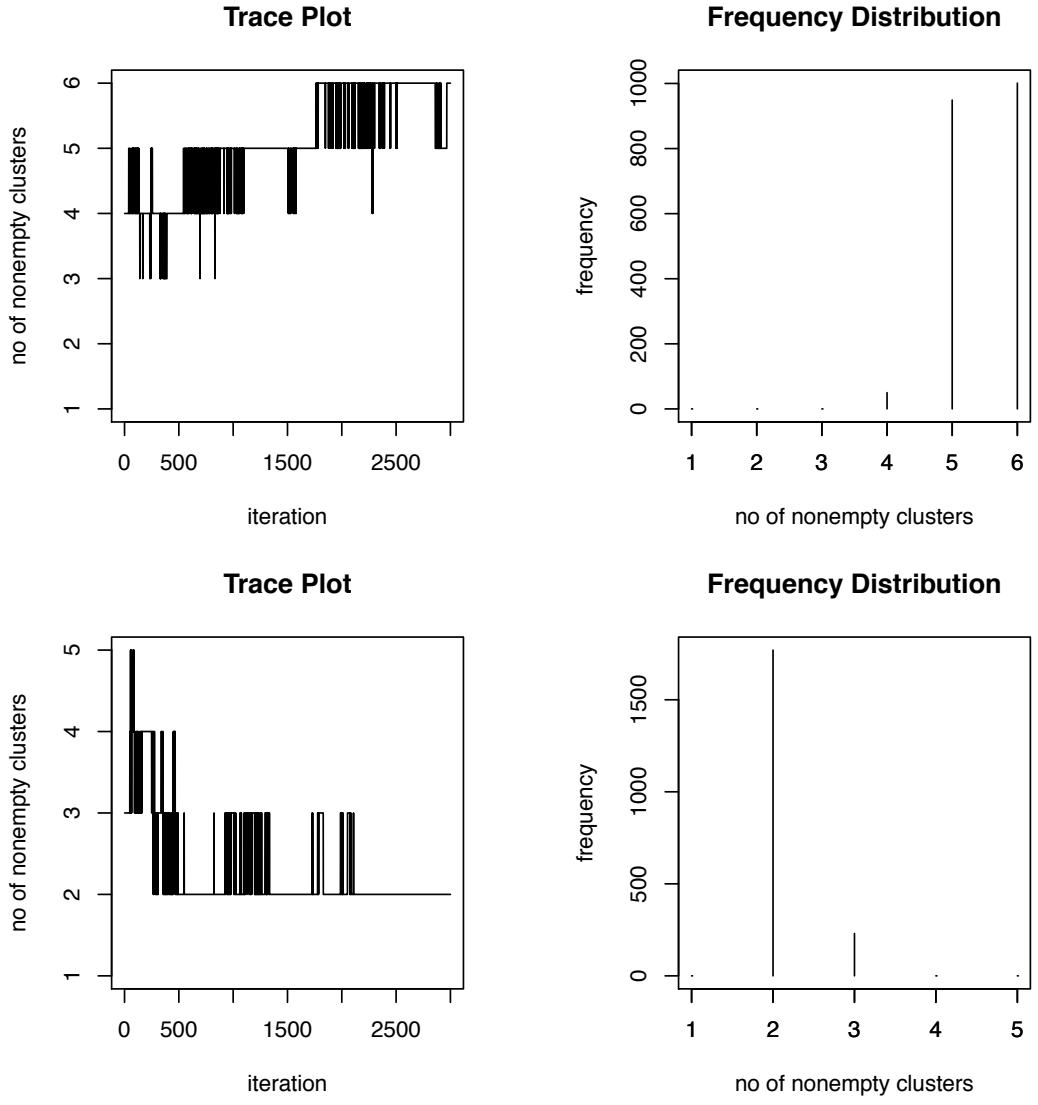

Figure 4.6: Trace plots and frequency distributions of the number of nonempty clusters produced by the MIW (mixtures with inverse Wishart priors) method for the conditionally heteroscedastic error distribution $f^2_{\boldsymbol{\epsilon}}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and identity matrix (I) for the component specific covariance matrices. See Section 4.8 for additional details. The results correspond to the simulation instance that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets, when the number of mixture components for both the density of interest and the density of scaled errors were kept fixed at $K_{\mathbf{X}} = 6$ and $K_{\boldsymbol{\epsilon}} = 5$. The upper panels are for the density of interest $f_{\mathbf{X}}$ and the lower panels are for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$. The true number of mixture components were $K_{\mathbf{X}} = 3$ and $K_{\boldsymbol{\epsilon}} = 3$. As can be seen from Figure 4.4, a mixture model with 2 nonempty clusters can approximate the true density of the scaled errors well.

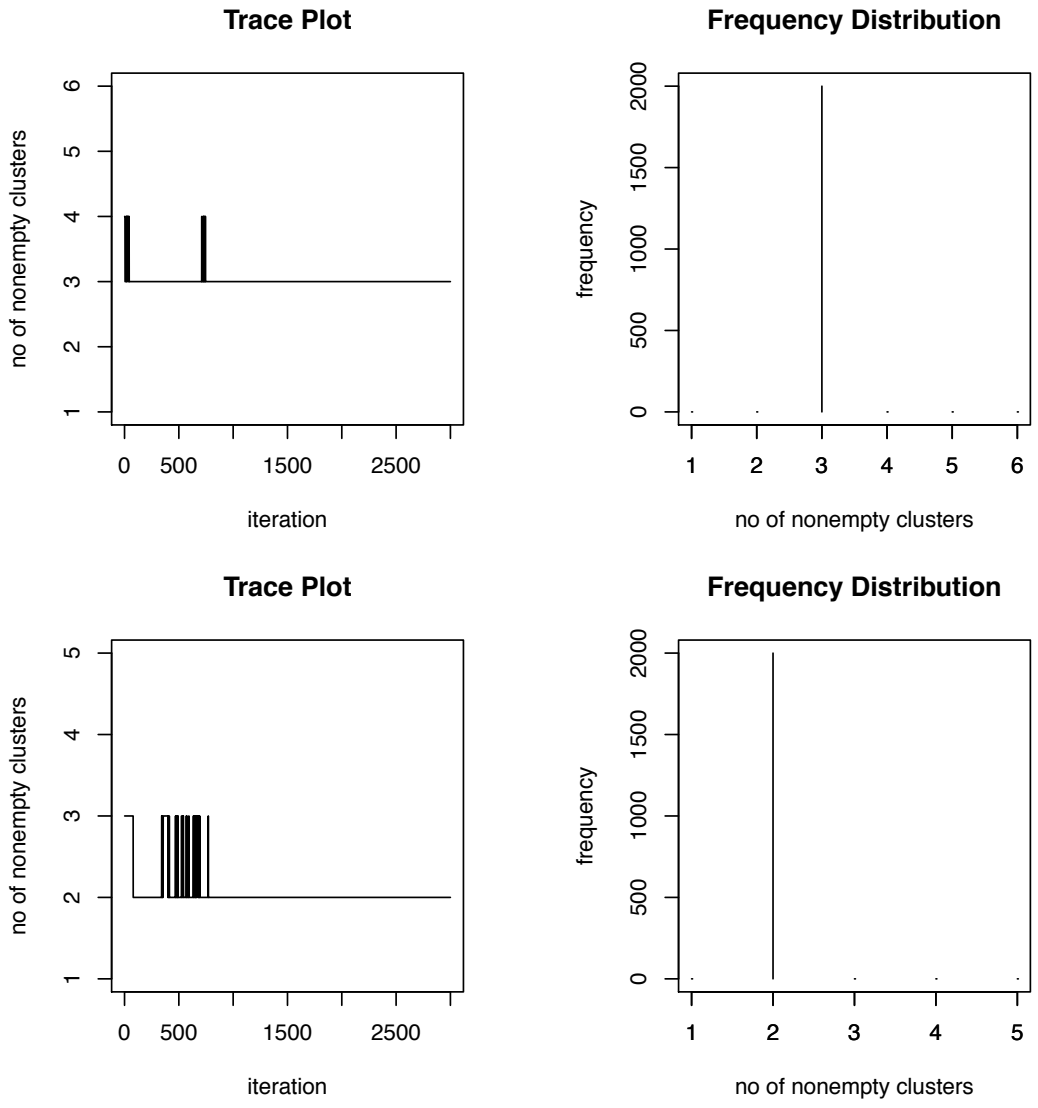

Figure 4.7: Trace plots and frequency distributions of the number of nonempty clusters produced by the MLFA (mixtures of latent factor analyzers) method for the conditionally heteroscedastic error distribution $f^2_{\boldsymbol{\epsilon}}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and identity matrix (I) for the component specific covariance matrices. See Section 4.8 for additional details. The results correspond to the simulation instance that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets, when the number of mixture components for both the density of interest and the density of scaled errors were kept fixed at $K_{\mathbf{X}} = 6$ and $K_{\boldsymbol{\epsilon}} = 5$. The upper panels are for the density of interest $f_{\mathbf{X}}$ and the lower panels are for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$. The true number of mixture components were $K_{\mathbf{X}} = 3$ and $K_{\boldsymbol{\epsilon}} = 3$. As can be seen from Figure 4.5, a mixture model with 2 nonempty clusters can approximate the true density of the scaled errors well.

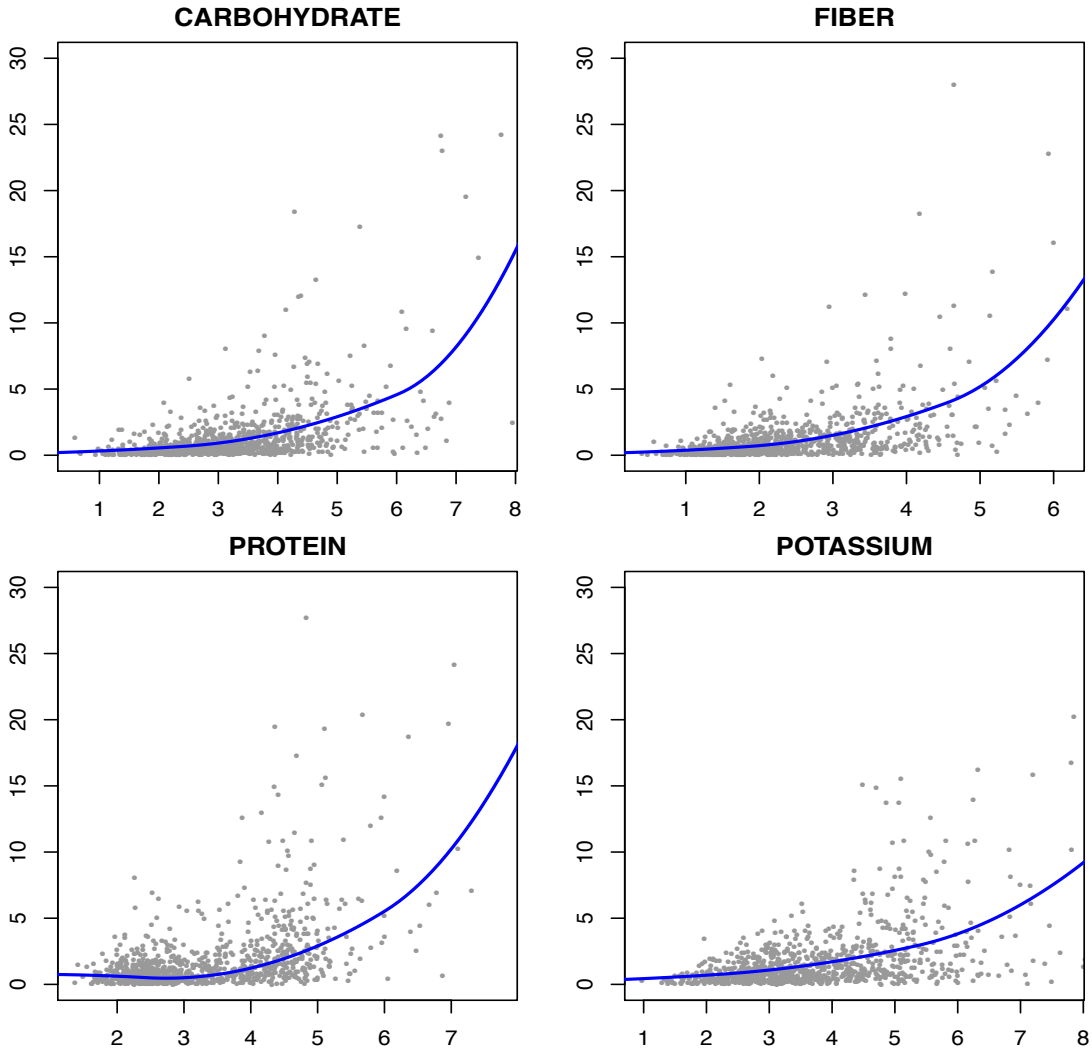

Figure 4.8: Estimated variance functions $s^2(X)$ produced by the univariate density deconvolution method for each component of the vector of interest $\mathbf{X}$ for the EATS data set with sample size $n = 965$, $m_i = 4$ replicates for each subject. See Section 4.9 for additional details. The figure is in color in the electronic version of this dissertation.

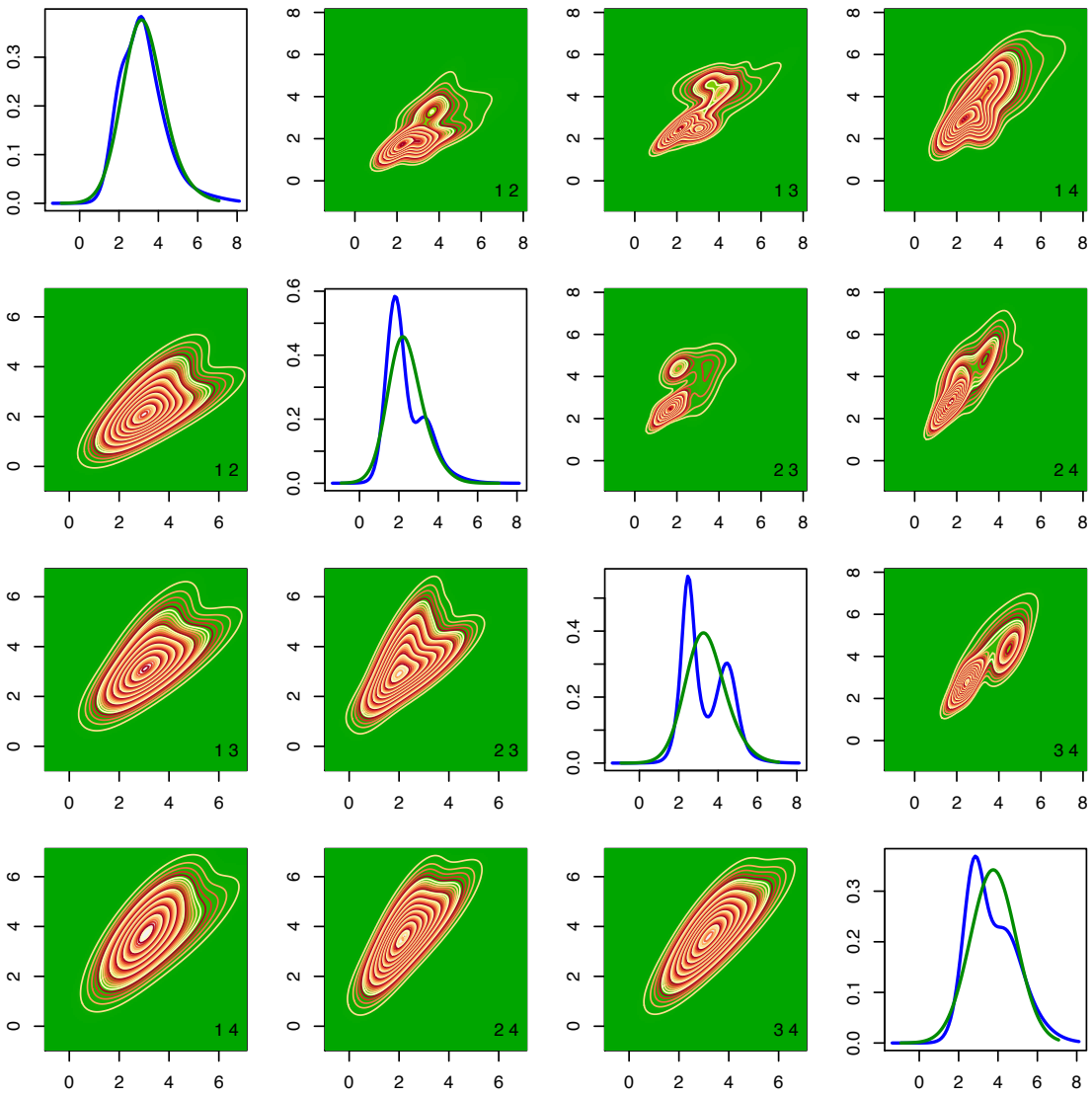

Figure 4.9: Results for the EATS data set for the density of interest $f_{\mathbf{X}}$. The off-diagonal panels show the contour plots of two-dimensional marginals estimated by the MIW method (upper triangular panels) and the MLFA method (lower triangular panels). The diagonal panels show the one dimensional marginal densities estimated by the MIW method (darker shaded lines) and the MLFA method (lighter shaded lines). The figure is in color in the electronic version of this dissertation.

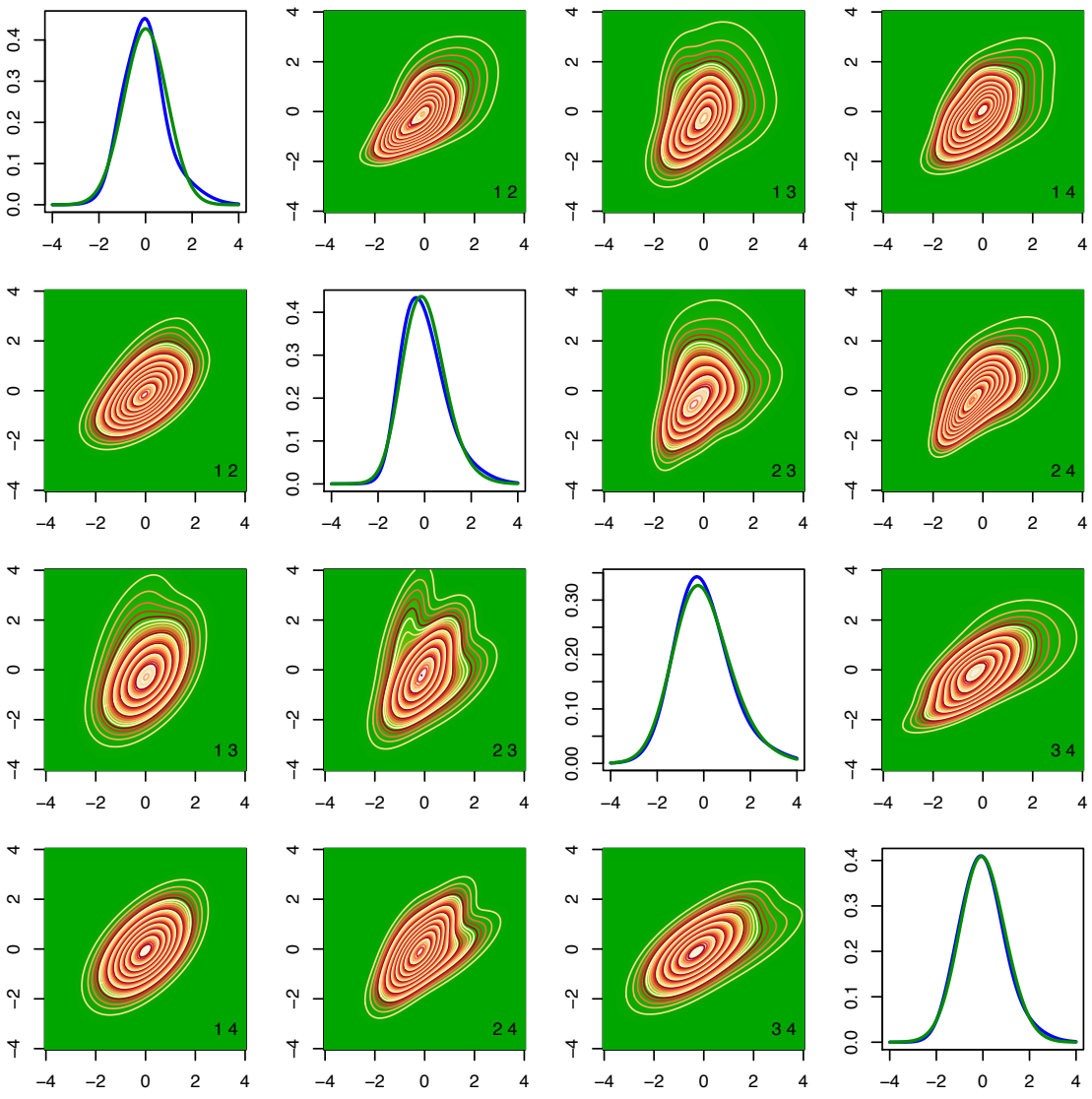

Figure 4.10: Results for the EATS data set for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$. The off-diagonal panels show the contour plots of two-dimensional marginals estimated by the MIW method (upper triangular panels) and the MLFA method (lower triangular panels). The diagonal panels show the one dimensional marginal densities estimated by the MIW method (darker shaded lines) and the MLFA method (lighter shaded lines). The figure is in color in the electronic version of this dissertation.

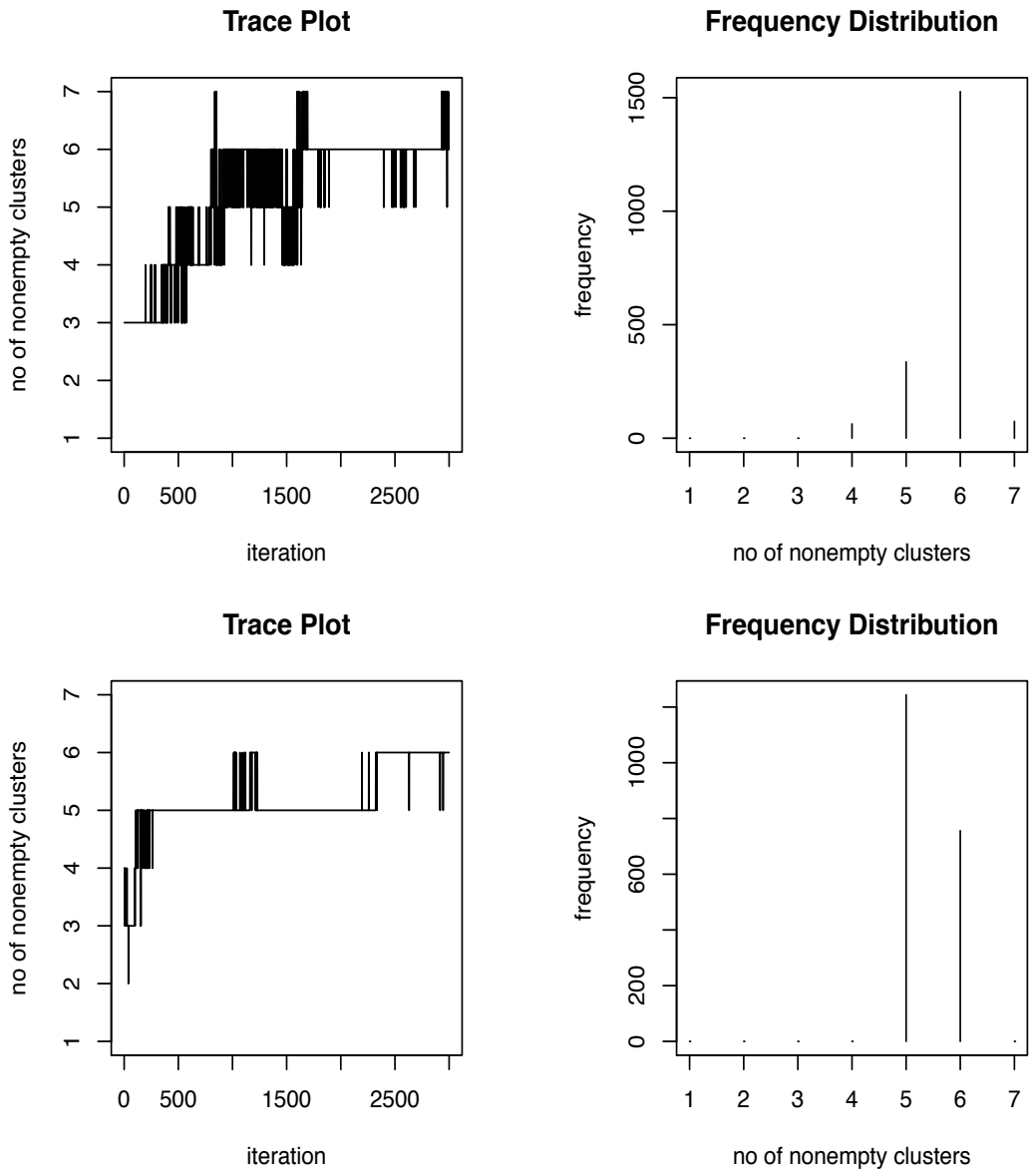

Figure 4.11: Trace plots and frequency distributions of the number of nonempty clusters produced by the MIW (mixtures with inverse Wishart priors) method for the EATS data example. See Section 4.9 for additional details. The number of mixture components for both the density of interest and the density of scaled errors were kept fixed at $K_{\mathbf{X}} = K_{\boldsymbol{\epsilon}} = 7$. The upper panels are for the density of interest $f_{\mathbf{X}}$ and the lower panels are for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$.

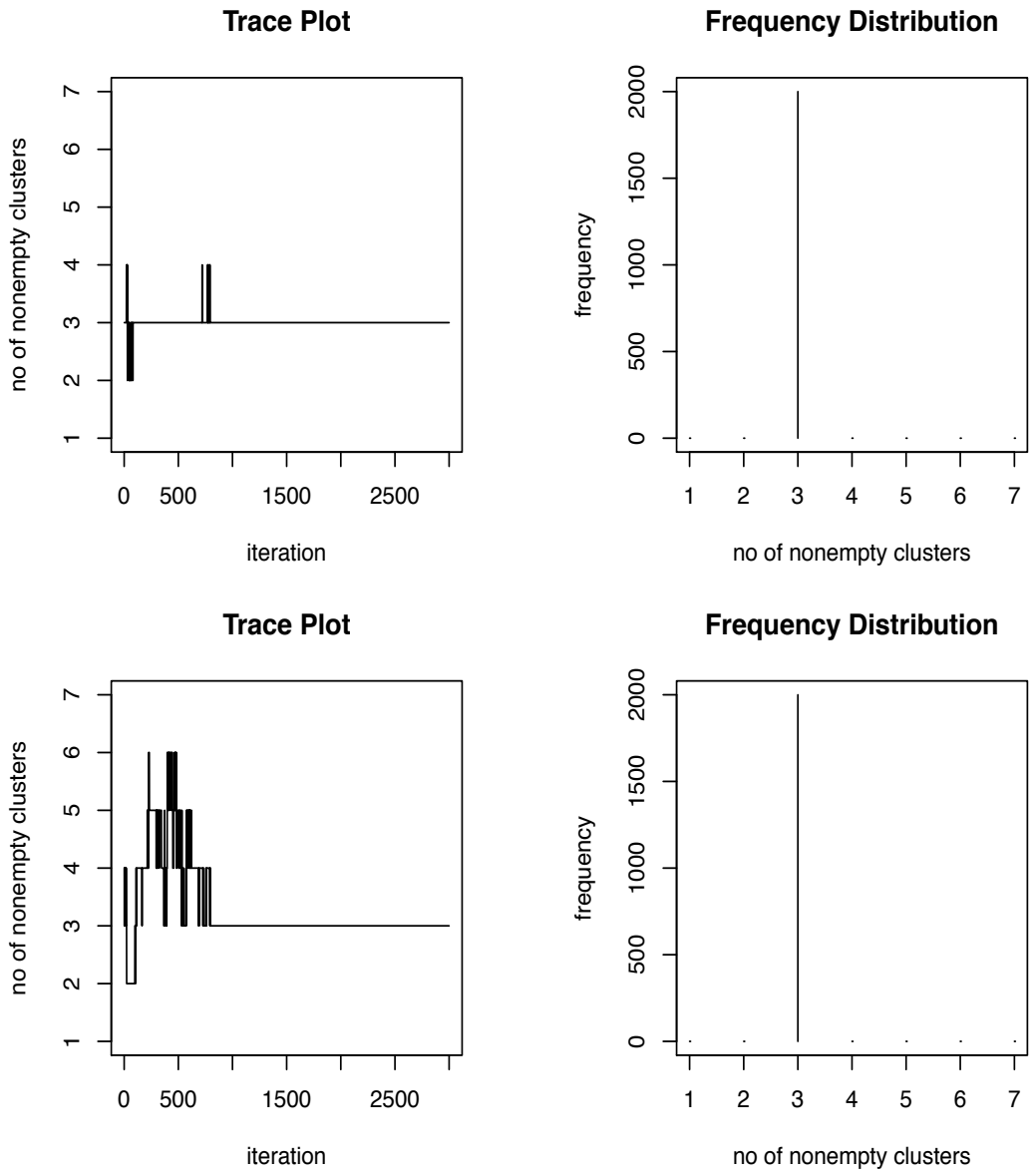

Figure 4.12: Trace plots and frequency distributions of the number of nonempty clusters produced by the MLFA (mixtures of latent factor analyzers) method for the EATS data example. See Section 4.9 for additional details. The number of mixture components for both the density of interest and the density of scaled errors were kept fixed at $K_{\mathbf{X}} = K_{\boldsymbol{\epsilon}} = 7$. The upper panels are for the density of interest $f_{\mathbf{X}}$ and the lower panels are for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$.

5. SUMMARY AND SCOPE FOR FUTURE RESEARCH

We first present a summary of the principal contributions of this dissertation. Throughout the course of the preceding sections, we developed a flexible Bayesian semiparametric framework for fundamentally important measurement error problems. The methods we developed relaxed many restrictive assumptions of previously existing techniques but also encompassed diverse simplified scenarios as special cases. These methods also provide the foundation for many interesting extensions and analyses, which we highlight following our summary of contributions.

5.1 Summary

Development of flexible and efficient Bayesian semiparametric methodology for important measurement error problems has been the primary focus of this dissertation. In previously existing literature, solutions to even the most fundamental measurement error problems like density deconvolution and regression with errors-in-predictors were available only under numerous simplifying and unrealistic assumptions. By accommodating measurement errors through natural hierarchies, we developed a very powerful Bayesian framework for solving these important measurement errors problems under less restricted and more realistic scenarios.

In Section 2, we developed univariate density deconvolution approaches when replicated proxies are available for each unknown value of the variable of interest and the variability of the measurement errors depends on the associated unobserved value of the variable of interest through an unknown relationship. We modeled the density of interest by a flexible location-scale mixture of normals induced by a Dirichlet process. We assumed that the measurement errors can be factored into zero mean 'scaled errors' that are independent of the variable of interest, and a variance function component that explains the conditional heteroscedasticity. This multiplicative structural assumption on the measurement errors was implicit in Staudenmayer, et al. (2008), where the scaled errors were assumed to come from a standard normal distribution. We considered a more flexible representation of the scaled errors, modeling its density by a Dirichlet process induced mixture with each component of the mixture being itself a two-component normal mixture with its mean restricted at zero. The variance function was modeled using flexible mixtures of B-splines. The proposed deconvolution approach thus used flexible Dirichlet process mixture models

twice, first to model the density of interest and second to model the density of the scaled errors, freeing them both from restrictive parametric assumptions, while at the same time accommodating conditional heteroscedasticity through the variance function.

In Section 3, we extended the methodology to the problem of robust estimation of the regression relationship between a response and a latent covariate. We considered scenarios when precise measurements on the covariate are not available but error-prone surrogates for the unobserved covariate are available for each sampled unit. For modeling conditionally heteroscedastic regression and measurement errors, we assumed, as in Section 2, that they can be factored into scaled errors that are independent of the covariate and variance function components that explain the conditional heteroscedasticity. The density of the covariate and the densities of the scaled errors were modeled using flexible mixture models induced by Dirichlet processes. The regression function and the variance functions were modeled using flexible mixtures of B-splines.

In Section 4, we considered the problem of multivariate density deconvolution where the variable of interest is vector valued. In sharp contrast to the univariate case, the literature on multivariate density deconvolution is extremely sparse and, to the best of our knowledge, all existing multivariate deconvolution approaches assume the density of the measurement errors to be completely known. We proposed robust Bayesian semiparametric multivariate deconvolution approaches when the measurement error density is not known but replicated proxies are available for each unobserved value of the random vector. Additionally, we also allowed the measurement errors to be conditionally heteroscedastic. The multivariate nature of the problem brought in new modeling challenges and computational obstacles that precluded straightforward extension of the univariate deconvolution approaches of Section 2. As in Section 2, we employed mixture models to approximate both the density of interest and the density of the measurement errors but instead of using infinite mixtures induced by Dirichlet processes, we used finite mixtures of multivariate normal kernels with symmetric Dirichlet priors on the mixture probabilities to model the multivariate densities. The use of finite mixtures with exchangeable priors enabled us to significantly reduce computational complexity while retaining essentially the same flexibility as that of Dirichlet process based infinite dimensional models. Using factor-analytic representation of the component specific covariance

matrices with sparsity inducing shrinkage priors on the factor loading matrices, we were able to build models that were flexible yet parsimonious and hence numerically stable. For the multivariate problem, we proposed a new strategy to enforce the mean zero restriction on the density of the measurement errors by exploiting the exchangeability of the symmetric Dirichlet prior and basic properties of multivariate normal kernels. This new strategy was much easier to implement and hence particularly suitable for high dimensional applications. To meet the computational challenges posed by the complicated likelihood function for conditionally heteroscedastic multivariate measurement errors, we designed a novel two-stage procedure that first estimates the variance functions using reparametrized versions of the corresponding univariate submodels, and then, in the second stage, estimates the remaining model parameters plugging in the estimates of the variance functions obtained in the first stage.

We provided theoretical results showing the flexibility of the proposed methods. We illustrated the efficiency of the proposed methods through extensive simulation experiments where the methods vastly outperformed their competitors. The practical usefulness of the proposed methods was illustrated through applications in nutritional epidemiology. The univariate and multivariate deconvolution methods were applied to estimate consumption patterns of different dietary components from contaminated 24 hour recalls. The regression techniques developed in Section 3 were applied to estimate the regression relationship between the intakes reported in food frequency questionnaires and the latent true dietary intakes, by treating the contaminated 24 hour recalls as unbiased surrogates for the latent covariate.

To conclude, in this dissertation we developed robust ad efficient Bayesian semiparametric approaches for fundamentally important measurement error problems. With their theoretically proven flexibility, their much superior empirical performance over current state-of-the-art methods, their ability to encompass simple parametric models as special subcases, and their practical usefulness illustrated through important real world examples, these methods, we believe, make important contributions to the literature on measurement error problems.

5.2 Scope for Future Research

We conclude this dissertation with brief discussions of some interesting ongoing projects and some open directions for future research.

5.2.1 Deconvolution with Excess and Hard Zeroes

In the nutritional epidemiology examples discussed in Section 2 and Section 4, the problems of estimating the densities of the true intakes of regularly consumed dietary components were considered. These components being consumed on a daily basis, their reported intakes are all positive and continuously measured. In contrast, for 'episodically consumed' dietary component, the reported intake may equal zero on a non-consumption day, or is positive on a day the component is consumed. A third type of dietary components may be referred to as 'never consumed' components, ones that are never consumed by some members of the population but are consumed by the rest, though not necessarily on a daily basis. Data on episodically and never consumed dietary components are zero-inflated data with measurement error, excess zeros caused by episodic consumption and hard zeros caused by never consumption.

The problem of excess zeroes has been addressed in the literature by Tooze, et al. (2006) and Kipnis, et al. (2009). The problem of hard zeros has been considered by Kipnis, et al. (2009) and by Keogh and White (2011), but only for a single variable. Zhang, et al. (2011) considered the more important problem of estimating the consumption pattern of an episodically consumed dietary component and the distribution of usual intake of energy jointly that enables nutritionists to compare the diets of individuals with very different energy intakes. These approaches are based on multiple non-linear transformations and approximations and strong parametric assumptions, and thus suffer from all the short-comings of transformation-retransformation models discussed in Appendix A.5.

Extending the deconvolution methods described in Section 2 and Section 4 of this dissertation to accommodate excess and hard zeroes would be an interesting direction for further research. We have made some progress towards this direction with promising preliminary results.

5.2.2 Study of Asymptotic Properties

An element absent from this dissertation is the study of the asymptotic properties of the proposed Bayesian procedures.

Posterior consistency and rates of convergence of Bayesian estimators in ordinary density estimation problems have been studied by Ghosal, et al. (1999), Ghosal, et al. (2000), Ghosal and van der Vaart (2007a, 2007b) and Shen and Wasserman (2001). Posterior consistency and rates convergence of Bayesian estimators in semi-

parametric regression problems with precisely measured covariates have been studied by Amewou-Atisso, et al. (2003) and Pelenis (2014), among others.

In measurement error problems, the asymptotic behavior of the posterior would be much more challenging to assess, and, to the best of our knowledge, have not been studied in the literature, not even for parametric models. Consistency of the proposed Bayesian methods is not theoretically investigated in this dissertation either, but since the flexibility of the priors is intimately related to the consistency of the posterior estimates, the results of Appendix D provide a crucial first step in that direction. The MISE performances of the proposed models also provide empirical evidence in favor of consistency. We are studying convergence properties of Bayesian deconvolution models in simpler known measurement error density set up as the subject of separate research. The study of the asymptotic properties of the posterior in regression problems with errors in covariates would be an interesting research problem to explore next.

5.2.3 Flexible Covariance Regression Models

The technique proposed in Section 4.4.1 of this dissertation to enforce mean zero moment restriction on multivariate measurement errors is simple but quite powerful and its applications are not limited to deconvolution problems only. As commented upon in Section 4.4.2 and detailed in Appendix C, the covariance regression techniques developed in Hoff and Niu (2012) and Fox and Dunson (2013) for modeling conditionally varying regression errors are not suitable for modeling conditionally heteroscedastic measurement errors. Similarly the models for conditionally heteroscedastic multivariate measurement errors developed in this dissertation are not suitable for modeling conditionally heteroscedastic regression errors. However, the techniques proposed in Section 4.4.1 to enforce the mean zero moment restriction on multivariate measurement errors can be very easily adapted to relax the strong assumption of multivariate normality of multivariate regression errors made in both Hoff and Niu (2012) and Fox and Dunson (2013).

5.2.4 Development of Sophisticated Software Packages

The methods developed in this dissertation were all implemented by programs written in R. Since our methods are based on Bayesian hierarchical framework that include many parametric models as special subcases, these special subcases can be easily accommodated in the codes using simple binary switches. For example, for

the univariate density deconvolution models described in Section 2, the parameters specifying the density MCMC chain were initialized at values that correspond to normally distributed errors. Density deconvolution with normally distributed measurement errors can thus be trivially implemented simply by switching off the updates for these parameters without having to write separate codes. In the same vain, the spacial case of homoscedastic measurement errors can be accommodated simply by switching of the roles of the variance functions that model the conditional variability.

The flexibility of our proposed methods, the ease with which they can accommodate numerous special subcases and the significant improvements they achieve over existing methods lead us to believe there is place for a new R package that would implement these methods making them accessible to a broader audience. There is, however, ample scope of improving the current versions of the R codes. It is certainly possible to improve the speed by translating at least segments of the codes to a more efficient low level programming languages, like C, C++ or JAVA. The starting values of the MCMC sampler implementing the multivariate deconvolution methods described in Section 4 were determined by first running the corresponding univariate submodels. Parallelizing these runs can also result in significant gain in computing time. We have started incorporating these improvements into our R prototypes.

REFERENCES

Amewou-Atisso, M., Ghosal, S., Ghosh, J. K. and Ramamoorthi, R.V. (2003). Posterior consistency for semi-parametric regression problems. *Bernoulli*, 9, 183-371.

Azzalini, A. (1985). A class of distributions which includes the Normal ones. *Scandinavian Journal of Statistics*, 12, 171-178.

Berry, S. A., Carroll, R. J. and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97, 160-169.

Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika*, 98, 291-306.

Böhning, D., Dietz, E. and Schlattmann, P. (1998). Recent developments in computer-assisted analysis of mixtures. *Biometrics*, 54, 367-377.

Bovy, J., Hogg, D. W. and Rowies, S. T. (2011). Extreme deconvolution: inferring complete distribution functions from noisy, heterogeneous and incomplete observations. *Annals of Applied Statistics*, 5, 1657-1677.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods and Applications*. New York: Chapman and Hall/CRC.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83, 1184-1186.

Carroll, R. J. and Hall, P. (2004). Low order approximations in deconvolution and regression with errors in variables. *Journal of the Royal Statistical Society, Series B*, 66, 31-46.

Carroll, R. J., Maca, J. D. and Ruppert, D. (1999a). Non-parametric regression in presence of measurement error. *Biometrika*, 86, 541-554.

Carroll, R. J., Roeder, K. and Wasserman, L. (1999b). Flexible parametric measurement error models. *Biometrics*, 55, 44-54.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models*, Second Edition. Boca Raton: Chapman

and Hall/CRC Press.

Chan, D., Kohn, R., Nott, D. and Kirby, C. (2006). Locally adaptive semiparametric estimation of the mean and variance functions in regression models. *Journal of Computational and Graphical Statistics*, 15, 915-936.

Cheng, C. L. and Riu, J. (2006). On estimating linear relationships when both variables are subject to heteroscedastic measurement errors. *Technometrics*. 48, 511-519.

Chung, Y. and Dunson, D.B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104, 1646-1660.

Comte, F. and Lacour, C. (2013). Anisotropic adaptive density deconvolution. *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques*, 49, 569-609.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89, 1314-1328.

de Boor, C. (2000). *A Practical Guide to Splines*. New York: Springer.

Delaigle, A., Hall, P. and Meister, A. (2008). On deconvolution with repeated measurements. *Annals of Statistics*, 36, 665-685.

Delaigle, A. and Meister, M. (2007). Nonparametric regression estimation in the heteroskedastic errors-in-variables problem. *Journal of the American Statistical Association*, 102, 1416-1426.

Delaigle, A. and Meister, M. (2008). Density estimation with heteroscedastic error. *Bernoulli*, 14, 562-579.

DerSimonian, R. (1986). Algorithm AS 221: maximum likelihood estimation of a mixing distribution. *Applied Statistics*, 35, 302-309.

Devanarayan, V. and Stefanski, L. A. (2002). Empirical simulation extrapolation for measurement error models with replicate measurements. *Statistics and Probability Letters*, 59, 219-225.

Devroye, L. (1989). Consistent deconvolution in density estimation. *Canadian Journal of Statistics*, 17, 235-239.

Diggle, P. J. and Hall, P. (1993). A Fourier approach to nonparametric deconvolution of a density estimate. *Journal of the Royal Statistical Society, Series B*, 55, 523-531.

Eckert, R. S., Carroll, R. J. and Wang, N. (1997). Transformations to additivity in measurement error models. *Biometrics*, 53, 262-272.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89-121.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90, 577-588.

Fan, J. (1991a). On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, 19, 1257-1272.

Fan, J. (1991b). Global behavior of deconvolution kernel estimators. *Statistica Sinica*, 1, 541-551.

Fan, J. (1992). Deconvolution with supersmooth distributions. *Canadian Journal of Statistics*, 20, 155-169.

Fan, J. and Koo, (2002). Wavelet deconvolution. *IEEE Transactions on Information Theory*, 48, 734-747.

Fan, J. and Truong, Y. (1993). Nonparametric regression with errors in variables. *Annals of Statistics*, 21, 1900-1925.

Ferguson, T. F. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 209-230.

Fokoué, E. and Titterington, D. M. (2003). Mixtures of factor analyzers. Bayesian estimation and inference by stochastic simulation. *Machine Learning*, 50, 73-94.

Fox, E. B. and Dunson, D. (2013). Bayesian nonparametric covariance regression. *Arxiv preprint arXiv:1101.2017*.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer.

Fuller, W. A. (1987). *Measurement Error Models*. New York: John Wiley & Sons.

Geweke, J. (2007). Interpretation and inference in mixture models: Simple MCMC works. *Computational Statistics & Data Analysis*, 51, 3529-3550.

Ghosal, S., Ghosh, J. K. and Ramamoorthy, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics*, 27, 143-158.

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, 28, 500-531.

Ghosal, S. and van der Vaart, A. W. (2007a). Convergence rates of posterior distributions for non-iid observations. *Annals of Statistics*, 35, 192-223.

Ghosal, S. and van der Vaart, A. W. (2007b). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Annals of Statistics*, 35, 697-723.

Ghosh, J. K. and Ramamoorthi, R. V. (2010). *Bayesian Nonparametrics*. New York: Springer.

Hastie, D. I., Liverani, S. and Richrdson, S. (2013). Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations. *Arxiv preprint arXiv:1304.1778*.

Hesse, C. H. (1999). Data driven deconvolution. *Journal of Nonparametric Statistics*, 10, 343-373.

Hoff, P. D. and Niu, X. (2012). A covariance regression model. *Statistica Sinica*, 22, 729-753.

Hu, Y. and Schennach, S. M. (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76, 195-216.

Ishwaran, H. and Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87, 371-390.

Ishwaran, H. and Zarepour, M. (2002). Exact and approximate sum-representations for the Dirichlet process. *Canadian Journal of Statistics*, 30, 269-283.

Keogh, R. H. and White, I. R. (2011). Allowing for never and episodic consumers when correcting for error in food record measurements of dietary intake. *Biostatistics*, 12, 624-636.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 27, 887-906.

Kipnis, V., Midthune, D., Buckman, D. W., Dodd, K. W., Guenther, P. M., Krebs-Smith, S. M., Subar, A. F., Tooze, J. A., Carroll, R. J. and Freedman, L. S. (2009). Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics*, 65, 1003-1010.

Li, T. and Vuong, Q. (1998). Nonparametric estimation of the measurement error model using multiple indicators. *Journal of Multivariate Analysis*, 65, 139-165.

Liard, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73, 805-811.

Lindsay, B. G. (1983a). Geometry of mixture likelihoods: a general theory. *Annals of Statistics*, 11, 86-94.

Lindsay, B. G. (1983b). Geometry of mixture likelihoods, part II: the exponential family. *Annals of Statistics*, 11, 783-792.

Liu, M. C. and Taylor, R. L. (1989). A consistent nonparametric density estimator for the deconvolution problem. *Canadian Journal of Statistics*, 17, 427-438.

Liu, A., Tong, T. and Wang, Y. (2007). Smoothing spline estimation of variance functions. *Journal of Computational and Graphical Statistics*, 16, 312-329.

Masry, E. (1991). Multivariate probability density deconvolution for stationary random processes. *IEEE Transactions on Information Theory*, 37, 1105-1115.

McIntyre, J. and Stefanski, L. A. (2011). Density estimation with replicate heteroscedastic measurements. *Annals of the Institute of Statistical Mathematics*, 63, 81-99.

Mengersen, K. L., Robert, C. P. and Titterington, D. M. (eds) (2011). *Mixtures - Estimation and Applications*. New York: John Wiley & Sons.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9, 249-265.

Neumann, M. H. (1997). On the effect of estimating the error density in nonparametric deconvolution. *Journal of Nonparametric Statistics*, 7, 307-330.

Norets, A. and Pelenis, J. (2012). Bayesian modeling of joint and conditional distributions. *Journal of Econometrics*, 168, 332-346.

Pati, D. and Dunson, D. (2013). Bayesian nonparametric regression with varying residual density. *Annals of the Institute of Statistical Mathematics*, 66, 1-13.

Pelenis, J. (2014). Bayesian regression with heteroscedastic error density and parametric mean function. *Journal of Econometrics*, 178, 624-638.

Pensky, M. and Vidakovic, B. (1999). Adaptive wavelet estimator for nonparametric density deconvolution. *Annals of Statistics*, 27, 2033-2053.

Pilla, R. S. and Lindsay, B. G. (2001). Alternative EM methods for nonparametric finite mixture models. *Biometrika*, 88, 535-555.

Rasmussen and Williams. (2006). *Gaussian Processes for Machine Learning*. Cambridge: MIT Press.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behavior of the posterior distribution in overfitted mixture models *Journal of the Royal Statistical Society, Series B*, 73, 689-710.

Sarkar, A., Mallick, B. K. and Carroll, R. J. (2014). Bayesian semiparametric regression in the presence of conditionally heteroscedastic measurement and regression errors. Accepted for publication in *Biometrics*.

Sarkar, A., Mallick, B. K., Staudenmayer, J., Pati, D. and Carroll, R. J. (2014). Bayesian semiparametric density deconvolution in the presence of conditionally heteroscedastic measurement errors. Forthcoming in *Journal of Computational and Graphical Statistics*. DOI: 10.1080/10618600.2014.899237.

Schennach, S. M. (2004a). Estimation of nonlinear models with measurement error. *Econometrica*, 72, 33-75.

Schennach, S. M. (2004b). Nonparametric regression in the presence of measurement error. *Econometric Theory*, 20, 1046-1093.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4, 639-650.

Shen, W. and Wasserman, L. (2001). Rates of convergence of posterior distribution. *Annals of Statistics*, 29, 687-714.

Spiegelman, D., Carroll, R. J. and Kipnis, V. (2001). Efficient regression calibration for logistic regression in main study/internal validation study designs with an

imperfect reference instrument. *Statistics in Medicine*, 20, 139-160.

Spiegelman, D., McDermott, A. and Rosner, B. (1997). The regression calibration method for correcting measurement error bias in nutritional epidemiology. *American Journal of Clinical Nutrition*, 65 (supplement), 1179S-1186S.

Spiegelman, D., Zhao, B. and Kim, J. (2005). Correlated errors in biased surrogates: study designs and methods for measurement error correction. *Statistics in Medicine*, 24, 1657-1682.

Staudenmayer, J., Ruppert, D. and Buonaccorsi, J. P. (2008). Density estimation in the presence of heteroscedastic measurement error. *Journal of the American Statistical Association*, 103, 726-736.

Stefanski, L. A. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics*, 21, 169-184.

Subar, A. F., Thompson, F. E., Kipnis, V., Midthune, D., Hurwitz, P. McNutt, S., McIntosh, A. and Rosenfeld, S. (2001). Comparative validation of the block, Willet, and National Cancer Institute food frequency questionnaires. *American Journal of Epidemiology*, 154, 1089-1099.

Tokdar, S. T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics*, 67, 90-110.

Tooze, J. A., Midthune, D., Dodd, K. W., Freedman, L. S., Krebs-Smith, S. M., Subar, A. F., Guenther, P. M., Carroll, R. J. and Kipnis, V. (2006). A new statistical method for estimating the usual intake of episodically consumed foods with application to their distribution. *Journal of the American Dietetic Association*, 106, 1575-1587.

Wang, X. and Wang, B. (2010). Estimating smooth distribution function in the presence of heteroscedastic measurement errors. *Computational Statistics & Data Analysis*, 54, 25-36.

Wang, X. and Wang, B. (2011). Deconvolution estimation in measurement error models: the R package decon. *Journal of Statistical Software*, 39, 1-24.

West, M., Müller, P. and Escobar, M. D. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. *Aspects of uncer-*

tainty: a tribute to D. V. Lindley, eds. A. F. M. Smith and P. Freeman, New York: Wiley, 363-386.

Willett, W. (2012). *Nutritional Epidemiology*, Third Edition. New York: Oxford University Press.

Yau, P. and Kohn, R. (2003). Estimation and variable selection in nonparametric heteroscedastic regression. *Statistics and Computing*, 13, 191-208.

Youndjé, E. and Wells, M. T. (2008). Optimal bandwidth selection for multivariate kernel deconvolution. *TEST*, 17, 138-162.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H. and Weir, B. S. (2002). Truncated product method for combining p-values. *Genetic Epidemiology*, 22, 170-185.

Zhang, S. Midthune, D., Pérez, A, Buckman, D. W., Kipnis, V., Freedman, L. S., Dodd, K. W., Krebs-Smith, S. M. and Carroll, R. J. (2011). Fitting a bivariate measurement error model for episodically consumed dietary components. International *Journal of Biostatistics*, Volume 7, Issue 1, Article 1, DOI: 10.2202/1557-4679.1267.

APPENDIX A

APPENDIX TO SECTION 2

A.1 Model Identifiability

Hu and Schennach (2008) showed that models such as ours are identified under very weak conditions. They show that when four variables, (Y, W, Z, X), where X is the only unobserved variate, are continuously distributed, their joint distribution is identified under the following conditions; their conditions are even weaker, but these suffice for our case.

Conditions 1. *1.* $f_{Y|W,Z,X} = f_{Y|X}$. *2.* $f_{W|Z,X} = f_{W|X}$. *3.* $\mathbb{E}(W \mid X) = X$. *4. The set* $\{Y : f_{Y|X}(Y \mid X_1) \neq f_{Y|X}(Y \mid X_2)\}$ *has positive probability under the marginal of* Y *for all* $X_1 \neq X_2$. *5. The marginal, joint and conditional densities of* (Y, W, Z, X) *are bounded.*

They also have a highly technical assumption about injectivity of operators, which is satisfied if the distributions of W given X and Z given X are complete. This is a weak assumption. This means, for example, that if $\int g(W) f_{W|X}(W \mid X) dW = 0$ for all X, then $g \equiv 0$.

When $m_i \geq 3$, identifiability of our model (1)-(2) is assured as it falls within the general framework of Hu and Schennach (2008). To see this, replace their Y_i by our W_{i1}, their W_i by our W_{i2}, their Z_i by our W_{i3} and their X_i by our X_i. Conditions 3.1-3.4 then follow from the fact that $(\epsilon_{i1}, \epsilon_{i2}, \epsilon_{i3}, X_i)$ have a continuous distribution and are mutually independent with $E(\epsilon_{ij}) = 0$. Condition 3.5 follows assuming the variance function v is continuous.

We conjecture that model (1)-(2) is identifiable even with $m_i \geq 2$ under very weak assumptions. We have numerical evidence to support the claim.

A.2 Initial Values and Proposals for $\boldsymbol{\xi}$

The conditional posterior log-likelihood of $\boldsymbol{\xi}$ for Model-I is given by

$$\ell(\boldsymbol{\xi} \mid \sigma_\xi^2, \mathbf{X}_{1:n}) = -\frac{1}{2\sigma_\xi^2} \boldsymbol{\xi}^{\mathrm{T}} P \boldsymbol{\xi} - \sum_{i=1}^{n} \left\{ \frac{m_i}{2} \log v(X_i, \boldsymbol{\xi}) + \sum_{j=1}^{m_i} \frac{1}{2v(X_i, \boldsymbol{\xi})} (W_{ij} - X_i)^2 \right\}.$$

The initial values for the M-H sampler for $\boldsymbol{\xi}$ is obtained as $\boldsymbol{\xi}^{(0)} = \arg\max \ell(\boldsymbol{\xi} \mid 0.1, \overline{\mathbf{W}}_{1:n})$. Numerical optimization is performed using the optim routine in R with the analytical gradient supplied.

The covariance matrix of the random walk proposal for $\boldsymbol{\xi}$ is taken to be the inverse of the negative of the matrix of second partial derivatives of $\ell(\boldsymbol{\xi} \mid 0.1, \overline{\mathbf{W}}_{1:n})$ evaluated at $\boldsymbol{\xi}^{(0)}$. Expressions for the gradient and the second derivatives of $\ell = \ell(\boldsymbol{\xi} \mid \sigma_\xi^2, \mathbf{X}_{1:n})$ are given below.

$$
\begin{aligned}
\frac{\partial \ell}{\partial \xi_k} &= -\frac{(P\boldsymbol{\xi})_k}{\sigma_\xi^2} - \sum_{i=1}^{n} \left\{ m_i - \sum_{j=1}^{m_i} \frac{(W_{ij} - X_i)^2}{v(X_i, \boldsymbol{\xi})} \right\} \frac{b_{2,k}(X_i) \exp(\xi_k)}{2v(X_i, \boldsymbol{\xi})}, \\
\frac{\partial^2 \ell}{\partial \xi_k^2} &= -\frac{(P)_{kk}}{\sigma_\xi^2} - \sum_{i=1}^{n} \left\{ \sum_{j=1}^{m_i} \frac{(W_{ij} - X_i)^2}{v(X_i, \boldsymbol{\xi})} - \frac{m_i}{2} \right\} \frac{b_{2,k}(X_i)^2}{v(X_i, \boldsymbol{\xi})^2} \exp(2\xi_k) \\
&\quad - \sum_{i=1}^{n} \left\{ m_i - \sum_{j=1}^{m_i} \frac{(W_{ij} - X_i)^2}{v(X_i, \boldsymbol{\xi})} \right\} \frac{b_{2,k}(X_i) \exp(\xi_k)}{2v(X_i, \boldsymbol{\xi})}, \\
\frac{\partial^2 \ell}{\partial \xi_k \partial \xi_{k'}} &= -\frac{(P)_{kk'}}{\sigma_\xi^2} - \sum_{i=1}^{n} \left\{ \sum_{j=1}^{m_i} \frac{(W_{ij} - X_i)^2}{v(X_i, \boldsymbol{\xi})} - \frac{m_i}{2} \right\} \frac{b_{2,k}(X_i) b_{2,k'}(X_i)}{v(X_i, \boldsymbol{\xi})^2} \exp(\xi_k + \xi_{k'}).
\end{aligned}
$$

A.3 Quadratic B-splines

Consider knot-points $t_1 = t_2 = t_3 = A < t_4 < \cdots < B = t_{K+3} = t_{K+4} = t_{K+5}$, where $t_{3:(K+3)}$ are equidistant with $\delta = (t_4 - t_3)$. For $j = 3, 4, \ldots, (K+2)$, define

$$
b_{2,j}(X) = \begin{cases}
\{(X - t_j)/\delta\}^2/2 & \text{if } t_j \le X < t_{j+1}, \\
-\{(X - t_{j+1})/\delta\}^2 + (X - t_{j+1})/\delta + 1/2 & \text{if } t_{j+1} \le X < t_{j+2}, \\
\{1 - (X - t_{j+2})/\delta\}^2 & \text{if } t_{j+2} \le X < t_{j+3}, \\
0 & \text{otherwise.}
\end{cases}
$$

Also define

$$
b_{2,1}(X) = \begin{cases}
\{1 - (X - t_1)/\delta\}^2/2 & \text{if } t_3 \le X < t_4, \\
0 & \text{otherwise.}
\end{cases}
$$

$$
b_{2,2}(X) = \begin{cases}
-\{(X - t_3)/\delta\}^2 + (X - t_4)/\delta + 1/2 & \text{if } t_3 \le X < t_4, \\
\{1 - (X - t_4)/\delta\}^2/2 & \text{if } t_4 \le X < t_5, \\
0 & \text{otherwise.}
\end{cases}
$$

$$b_{2,K+1}(X) = \begin{cases} \{(X - t_{K+1})/\delta\}^2/2 & \text{if } t_{K+1} \leq X < t_{K+2}, \\ -\{(X - t_{K+2})/\delta\}^2 + (X - t_{K+2})/\delta + 1/2 & \text{if } t_{K+2} \leq X < t_{K+3}, \\ 0 & \text{otherwise.} \end{cases}$$

$$b_{2,K+2}(X) = \begin{cases} \{(X - t_{K+2})/\delta\}^2/2 & \text{if } t_{K+2} \leq X < t_{K+3}, \\ 0 & \text{otherwise.} \end{cases}$$

A.4 Additional Simulation Experiments

Here we present the results of additional simulation experiments when the true density of interest is a normalized mixture of B-splines: $f_X^3(X) \propto \sum_{k=1}^{7} b_{2,k}(X)c_k$ with $\mathbf{c} = (c_1, \ldots, c_7)^{\mathrm{T}} = (0, 0, 2, 0.1, 1, 0, 0)^{\mathrm{T}}$ and equidistant knots on $[-2, 6]$. The normalizing constant was estimated by numerical integration on a grid of 500 equidistant points in $[-2, 6]$. The true values of X were generated from f_X^3 using the inverse cumulative distribution function method. We recall that the SRB approach of Staudenmayer, et al. (2008) models f_X by normalized mixture of B-splines and assumes normality of the scaled errors. The SRB approach and the three methods we proposed in Section 2 are compared over a factorial combination of three sample sizes ($n = 250, 500, 1000$), nine different types of distributions for the scaled errors (Table 2.1 and Figure 2.3), and one variance function $v(X) = (1 + X/4)^2$. For each subject, $m_i = 3$ replicates were simulated. The estimated MISEs are presented in Table A.1. Results for error distribution (i) are summarized in Figure A.1.

The results show that the deconvolution approaches proposed in Section 3.2 outperform the SRB model in all 27 (3×9) cases, even in scenarios when the measurement errors were normally distributed and hence the truth actually conformed to the SRB model. This may be attributed to the fact that Models I, II and III estimate f_X by a flexible infinite mixture model, where the number of mixture components that are 'active' in the data is inferred semiautomatically from the data making it an adaptive data dependent approach. On the other hand, the SRB model estimates the density of interest by a mixture of normalized B-Splines with a fixed number of components. Model III, we recall, also relaxes parametric assumptions on the measurement errors, accommodating skewness, multimodality and heavy tails and resulting in huge reductions in MISE over other models when the measurement errors are heavy-tailed.

True Error Distribution	Sample Size	MISE ×1000			
		SRB	Model1	Model2	Model3
(a)	250	8.66	**4.58**	4.74	4.68
	500	4.80	**3.63**	3.74	3.87
	1000	4.03	**2.57**	2.75	2.68
(b)	250	9.13	5.77	**4.38**	4.48
	500	5.12	3.76	**3.53**	3.56
	1000	4.68	2.83	**2.50**	2.72
(c)	250	6.35	4.74	4.35	**4.16**
	500	6.08	3.15	3.85	**3.07**
	1000	3.93	2.54	2.96	**1.93**
(d)	250	6.31	5.17	5.95	**3.61**
	500	3.70	3.91	6.36	**2.70**
	1000	2.92	2.75	7.08	**2.03**
(e)	250	8.73	5.74	5.31	**4.06**
	500	7.42	5.63	3.70	**3.01**
	1000	7.99	3.37	2.35	**1.90**
(f)	250	8.86	5.32	5.39	**5.19**
	500	4.64	3.87	3.83	**3.12**
	1000	3.31	2.47	3.00	**2.35**
(g)	250	22.77	12.51	12.61	**3.45**
	500	19.66	17.66	17.09	**2.25**
	1000	40.55	22.66	16.36	**1.50**
(h)	250	11.15	6.61	6.38	**3.96**
	500	8.34	9.38	7.18	**3.22**
	1000	13.69	9.91	7.98	**2.03**
(i)	250	17.49	12.25	13.55	**3.28**
	500	32.99	20.40	15.19	**2.42**
	1000	40.67	19.47	12.18	**1.17**

Table A.1: Mean integrated squared error (MISE) performance of density deconvolution models described in Section 3.2 of this dissertation (Models I, II and III) compared with the model of Staudenmayer, et al. (2008) (Model SRB) for different scaled error distributions when the true density of interest is a mixture of splines. The true variance function was $v(X) = (1 + X/4)^2$. See Section A.4 for additional details. The minimum value in each row is highlighted.

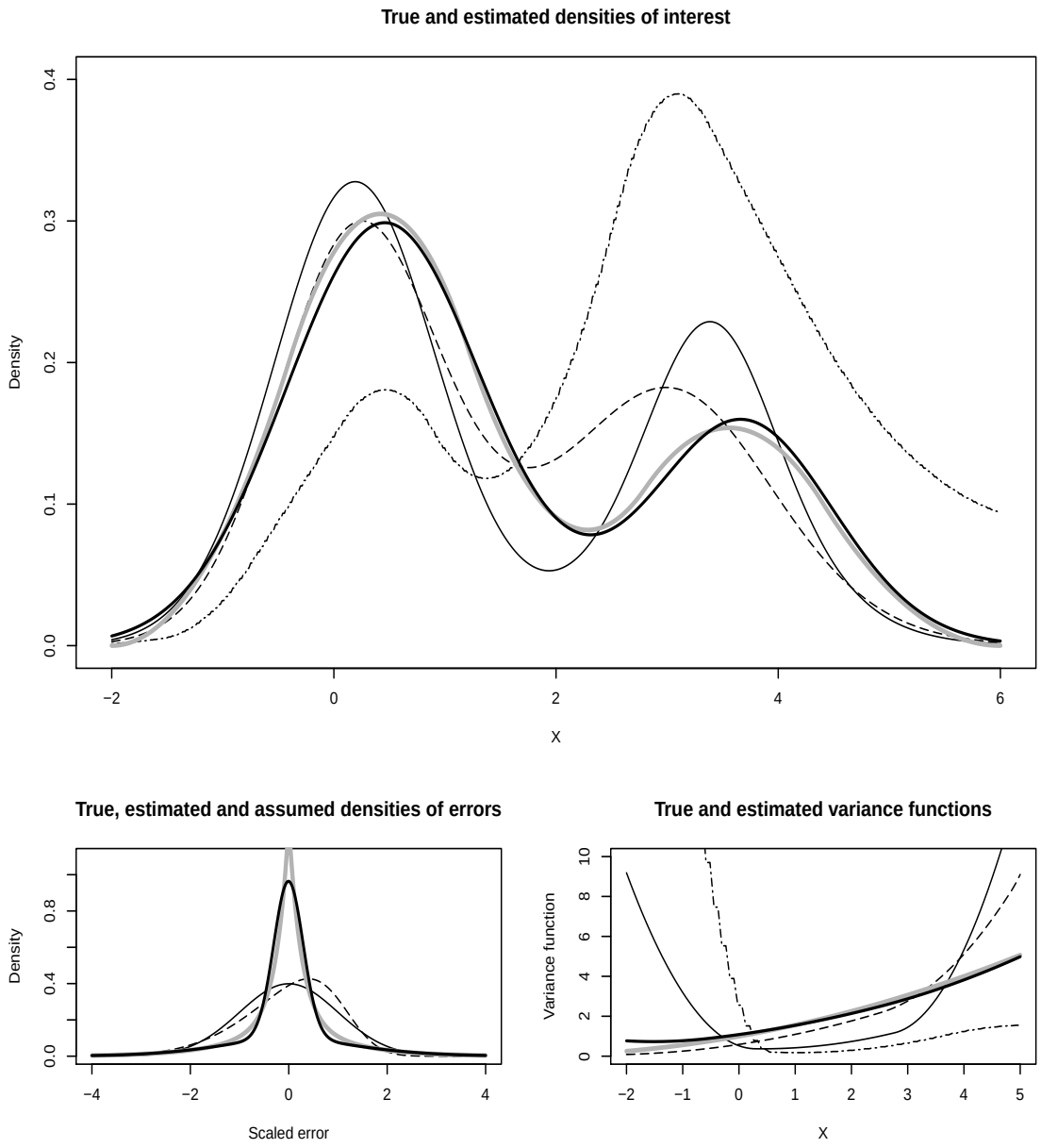

Figure A.1: Results for heavy-tailed error distribution (i) with sample size n=1000 corresponding to 25^{th} percentile MISE. The true density f_X is a normalized mixture of B-splines. See Section A.4 for additional details. The top panel shows the estimated densities under different models. The bottom left panel shows estimated densities of scaled errors under Model-II (dashed line) and Model-III (solid bold line) superimposed over a standard Normal density (solid line). The bottom right panel shows estimated variance functions under different models. For the top panel and the bottom right panel, the solid thin line is for Model-I; the dashed line is for Model-II; the solid bold line is for Model-III; and the dot-dashed line is for the Model of Staudenmayer, et al. (2008). In all three panels the bold gray lines represent the truth.

A.5 Transform-Retransform Methods

As described in the main text, in some problems interest is in the original scale while in other problems, interest is in the transformed scale. Here we describe transformations and transform-retransform methods.

In most cases, transformation to a scale that, hopefully, allows additivity and homoscedasticity of measurement errors is based on the Box-Cox family. The Box-Cox transformation of a variable $Z > 0$ is given by $h_\gamma(Z) = (Z^\gamma - 1)/\gamma$ if $\gamma \neq 0$ and $h_\gamma(Z) = \log(Z)$ if $\gamma = 0$. The Box-Cox transformation is applicable to strictly positive values only. It is common to estimate γ from the data and do analysis conditional on the estimated scale. In the presence of replicates there are at least three different ways to estimate γ. The most common practice is to estimate γ by maximizing the likelihood of the observed W_{ij} assuming the replicates W_{ij} to be normal and homoscedastic. A second alternative aimed at making the measurement errors normal and homoscedastic is to estimate γ by maximizing the p-value returned by standard tests of Normality of the differences $(W_{ij_1} - W_{ij_2})$. Yet another alternative is to estimate γ by minimizing the absolute correlation between the subject specific sample means $\overline{W}_i$ and variances S^2_{Wi}. aims to make the measurement errors independent of X. We refer to these three methods by R (for replicates), D (for differences), and I (for independence), respectively. There may not, however, exist a Box-Cox transformation (of either type) that can achieve even its primary goal satisfactorily well. Additionally, as noted in Carroll and Ruppert (1988), even if Box-Cox transformations exist that can separately remove non-normality, heteroscedasticity and dependence, there is no guarantee that a single transformation can do all.

If one wishes to estimate the density of the transformed $h_\gamma(X)$, and one assumes that $h_\gamma(W_{ij}) = h_\gamma(X_i) + U_{ij}$, then many methods are available, including ours. We will present simulations showing that our methods are generally much more efficient than frequentist competitors.

If one wishes to estimate the density of X on the original scale, then one must retransform. We have already remarked in Section 2 that retransformation violates the assumption that W is unbiased for X: it generally not possible to have unbiasedness on both the original and transformed scale. The reverse transformation is given by $g_\gamma(X) = (\gamma X + 1)^{1/\gamma}$ if $\gamma \neq 0$ and $g_\gamma(Z) = \exp(Z)$ if $\gamma = 0$. Let the subscript T denote the transformed scale. The model on the transformed scale is $W_T = X_T + U_T$, where for specificity here, we assume $U_T \sim \text{Normal}(0, \sigma^2_{U_T})$ and is independent of

X_T. Of course, our methodology does not make the normality assumption. To get the density of X, there are two methods. First, analytically, if the density of X_T is $f_{X_T}(x)$, then the density of X is $f_{X_T}\{h_\gamma(x)\}x^{\gamma-1}$. An alternative that we have found works just as well, and is somewhat more numerically stable, uses a second order Taylor series approximation,

$$\begin{aligned} W &= X + U = g_\gamma(W_T) = g_\gamma(X_T + U_T) \\ &\approx (\gamma X_T + 1)^{1/\gamma} + U_T\ (\gamma X_T + 1)^{1/\gamma - 1} + (1/2)\ U_T^2\ (1-\gamma)\ (\gamma X_T + 1)^{1/\gamma - 2} \\ &= (\gamma X_T + 1)^{1/\gamma} + (1/2)\ \sigma_{U_T}^2\ (1-\gamma)\ (\gamma X_T + 1)^{1/\gamma - 2} \\ &\quad + \{U_T\ (\gamma X_T + 1)^{1/\gamma - 1} + (1/2)\ (U_T^2 - \sigma_{U_T}^2)\ (1-\gamma)\ (\gamma X_T + 1)^{1/\gamma - 2}\}, \end{aligned}$$

where the adjustment in the last step ensures that the term inside the curly brackets is unbiased for zero. Using transform-retransform methods an estimate of the density on the observed scale f_X can be obtained from an estimate of f_{X_T} using the relationship $X \approx (\gamma X_T + 1)^{1/\gamma} + (1/2)\ \sigma_{U_T}^2\ (1-\gamma)\ (\gamma X_T + 1)^{1/\gamma - 2}$.

Most importantly, since the assumptions of normality, homoscedasticity and independence may still not be valid in the transformed scale and an approximate relationship is used to estimate f_X from $\widehat{f}_{X_T}$, the transform-retransform methods are often unable to capture interesting features of f_X such as multimodality etc.

A.6 Simulation Experiments

A.6.1 When Transformation to Additivity and Homoscedasticity Fails for Box-Cox Transformation

We first consider a setting that there is no Box-Cox transformation to exact additivity and homoscedasticity, and interest focuses on the original scale. The true density of interest is taken to be $f_X(x) = 0.8\ \text{Normal}(x \mid 1.5, 0.45^2) + 0.2\ \text{Normal}(x \mid 3, 0.30^2)\}$. This density assigns negligible probability to the negative axis. Two different densities are considered for the distribution of scaled errors: (a) Normal$(0, 1)$, and (b) an asymmetric bimodal density - density (c) of Section 2. The true variance function is taken to be $v(X) = \{(X/2)^2 + (X/3)^4\}$. For $n = 1000$ subjects, $m_i = m = 4$ replicates were simulated. The few replicates taking negative values were replaced by the corresponding true value of the variable of interest.

The performance of Model-III is compared with the performances of three different transform-retransform methods. The measurement error variance is estimated

by $\widehat{\sigma}^2_{U_T} = \sum_{ij}(W_{T,ij} - \overline{W}_{T,i})^2/\{n(m-1)\}$. The Bayesian independent error method (BIET) models the density in the transformed scale f_{X_T} by DPMM. The other two transform-retransform methods use the second order TAYLEX estimator of Carroll and Hall (2004) (CHT), including their bandwidth selector based on EBBS algorithm of Ruppert (1997), and the deconvoluting kernel estimator of Stefanski and Carroll (1990) and Carroll and Hall (1988) (DKET) to estimate f_{X_T}. For the DKET method the subject specific sample means $\overline{W}_{T,i}$ are used as available data with $\widehat{\sigma}^2_{U_T}/m$ as the measurement error variance. The DKET method is implemented by the DeconPdf function from the "decon" package in R (Wang and Wang, 2011). The bandwidth is selected by the "bw.dboot1" function from the "decon" package.

True Error Distribution	MISE ×1000									
	transform-retransform									Model 3 original scale
	BIET			CHT			DKET			
	D	I	R	D	I	R	D	I	R	
Normal	65.28	52.16	58.71	27.17	28.05	25.56	22.37	21.93	21.43	**4.24**
Mixture	51.61	41.67	53.58	42.32	43.46	40.21	22.44	21.29	24.62	**3.38**

Table A.2: MISE performance of Model III compared with the Bayesian independent error method (BIET), the Carroll and Hall Taylex method (CHT) and the deconvoluting kernel method (DKET) for two different measurement error distributions and for three different types of Box-Cox transformation applied to the replicates. The minimum value in each row is highlighted. Here "D" means the transformation that attempts to make the differences normally distributed, "I" attempts to make the errors homoscedastic and "R" aims to make the observed $h_\gamma(W_{ij})$ normally distributed.

The estimated MISEs are presented in Table A.2. Model-III vastly outperforms the transform-retransform methods. Figures A.2, A.3, A.4, A.5 illustrate in detail the implementation and the performances of the competing methods for normally distributed scaled errors. The true density is bimodal. The maximum p-value attained by the Box-Cox transformation that tries to to make the differences of the replicates normal by maximizing the p-value of Shapiro-Wilk test for normality is less than 10^{-16}. This shows that there does not exist a Box-Cox transformation that can make the measurement errors normal on the transformed scale. From the QQ plots in the third column of the rows B, C and D, it can be seen that none of the

three types of Box-Cox transformations can make the differences of the replicates (or equivalently the measurement errors) normal on the transformed scale. As can be seen from the fifth column of the rows B, C and D, the densities estimated by the transform-retransform methods are invariably unimodal. On the other hand, Model-III, that operates on the original scale and does not make restrictive assumptions, picks up the bimodality and estimates f_X quite well. Results when the scaled errors are skew normally distributed were very similar and are not presented. These examples illustrate the limitations of the transform-retransform methods discussed in Section 3.5 and reiterate the need for sophisticated deconvolution methods like Model-III introduced in Section 2 that can accommodate conditional heteroscedasticity and departures from normality and can operate directly on the observed scale.

We repeated the normal error distribution case with $n = 200$. The MISE efficiency for our method compared to the best transform-retransform method was 3.87, while it is 5.05 when $n = 1000$.

A.6.2 When Transformation to Additivity and Homoscedasticity is Possible

A.6.2.1 Original Scale Estimation, Heteroscedasticity

We next did simulations where it was possible to transform to additivity, homoscedasticity and normality. In this case, we used the same mixture of normals model for X as in Section A.6.1. The data in the original scale were generated as $W_{ij} = X_i \exp(U_{ij} - \sigma_u^2/2)$, where $U_{ij} \sim \text{Normal}(0, \sigma_u^2 = 0.135)$. We have unbiasedness in the original scale: $E(W_{ij}|X_i) = X_i$. The log-transformation gives $W_{ijT} = X_{iT} + U_{ijT} - \sigma_u^2/2$. We made the situation even more favorable for the transform-retransform method by assuming it was known that the log transformation was exact. We accounted for the term $\sigma_u^2/2$ in the retransformation. With $n = 200$, the MISE efficiency of our method to the best kernel transform-retransform method was 1.86, while for $n = 1000$ is was 2.23.

A.6.2.2 Transformed Scale Estimation, Homoscedasticity

Here we repeat the simulation of Section A.6.2.1 but define $X_{iT} = \log(X_i) - \sigma_u^2/2$, in other words, this is the classical deconvolution method. For $n = 200$ and $n = 1000$, the MISE efficiency of our methodology assuming homoscedasticity, referred to as the M3H model, compared to the best kernel method was 3.20 and 2.98, respectively. When we used our method but also estimated the (constant) variance function, the MISE efficiency of our methodology compared to the best kernel method was 2.41

and 2.92, respectively.

Sample Size	MISE ×1000			
	CH	DKE	Model 3	M3H
200	48.37	17.35	7.21	**5.41**
1000	30.69	6.24	2.14	**2.09**

Table A.3: MISE performance of Model III and Model III assuming homoscedasticity (M3H) compared with the Carroll and Hall Taylex method (CH) and the deconvoluting kernel method (DKE) for two sample sizes, all applied on the transformed scale. The minimum value in each row is highlighted.

A.7 Nutritional Epidemiology Example

Results for the daily intake of folate from the EATS data set were discussed in Section 3.4. Here we take the opportunity to discuss the results for the daily intake of a different dietary component, namely vitamin B, from the EATS data set. Figures A.6, A.7, A.8, A.9 summarize the results for Model-III and the transform-retransform methods. The maximum p-value attained by the Box-Cox transformation that tries to to make the differences of the replicates normal using the Shapiro-Wilk test for normality is less than 10^{-5}. As in the simulated examples, the QQ plots in the second and the third columns of the rows B, C and D also indicate that there does not exist a Box-Cox transformation that can make the replicates or the differences of the replicates normal in the transformed scale. The density estimated by Model-III is bimodal, whereas the densities estimated by the transform-retransform methods are all unimodal, similar to what is seen in Section A.6.1.

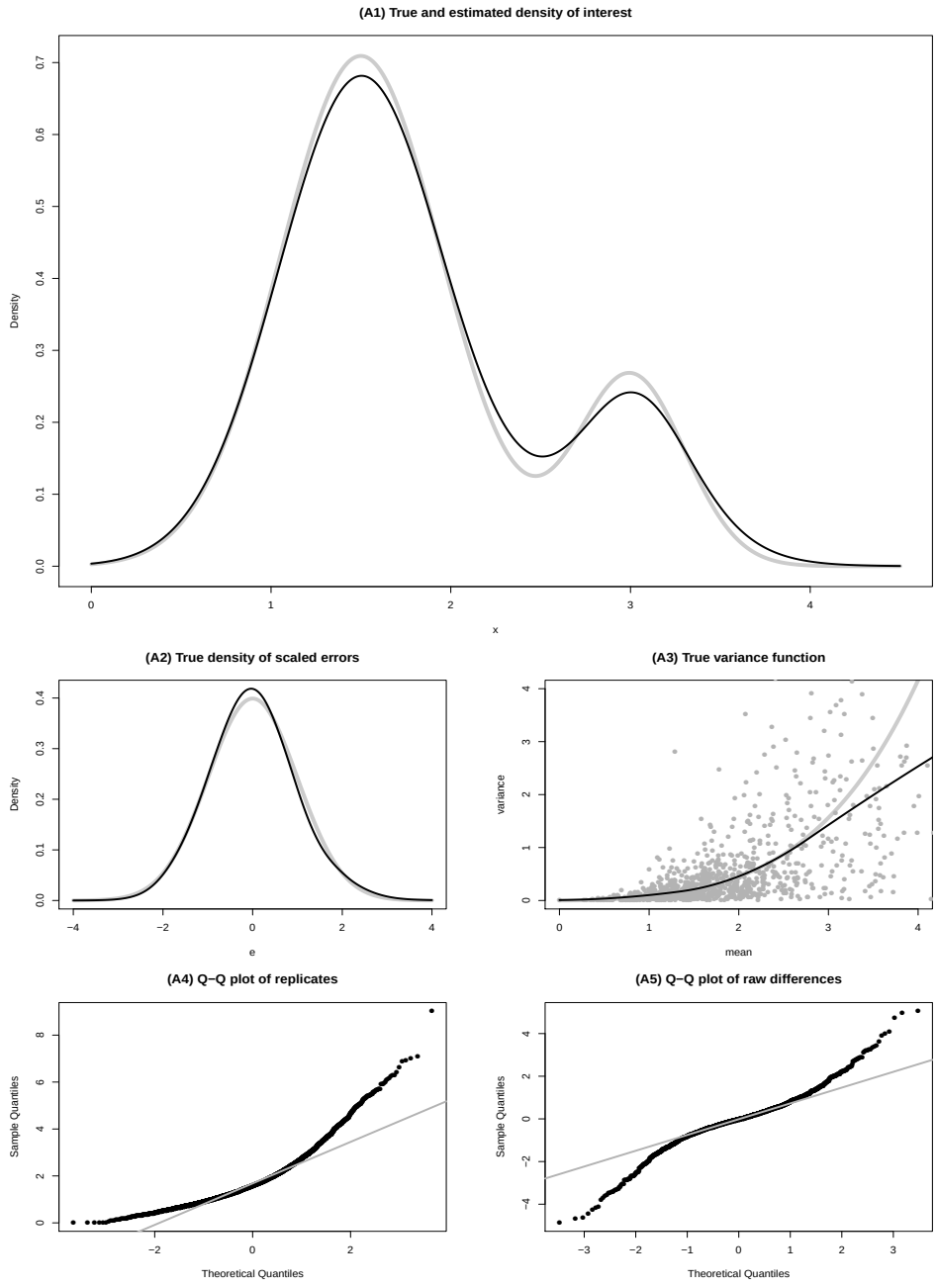

Figure A.2: Simulation results that illustrate the performance of Model-III in the original scale. Plot A1 shows the true (bold gray line) and estimated density of interest (bold black line); plot A2 shows the true (gray line) and the estimated density of measurement errors (bold black line); plot A3 shows the true and the estimated variance function superimposed with subject specific means and variances; plot A4 shows the Q-Q plot of the replicates; plot A5 shows the Q-Q plot of the differences of the replicates.

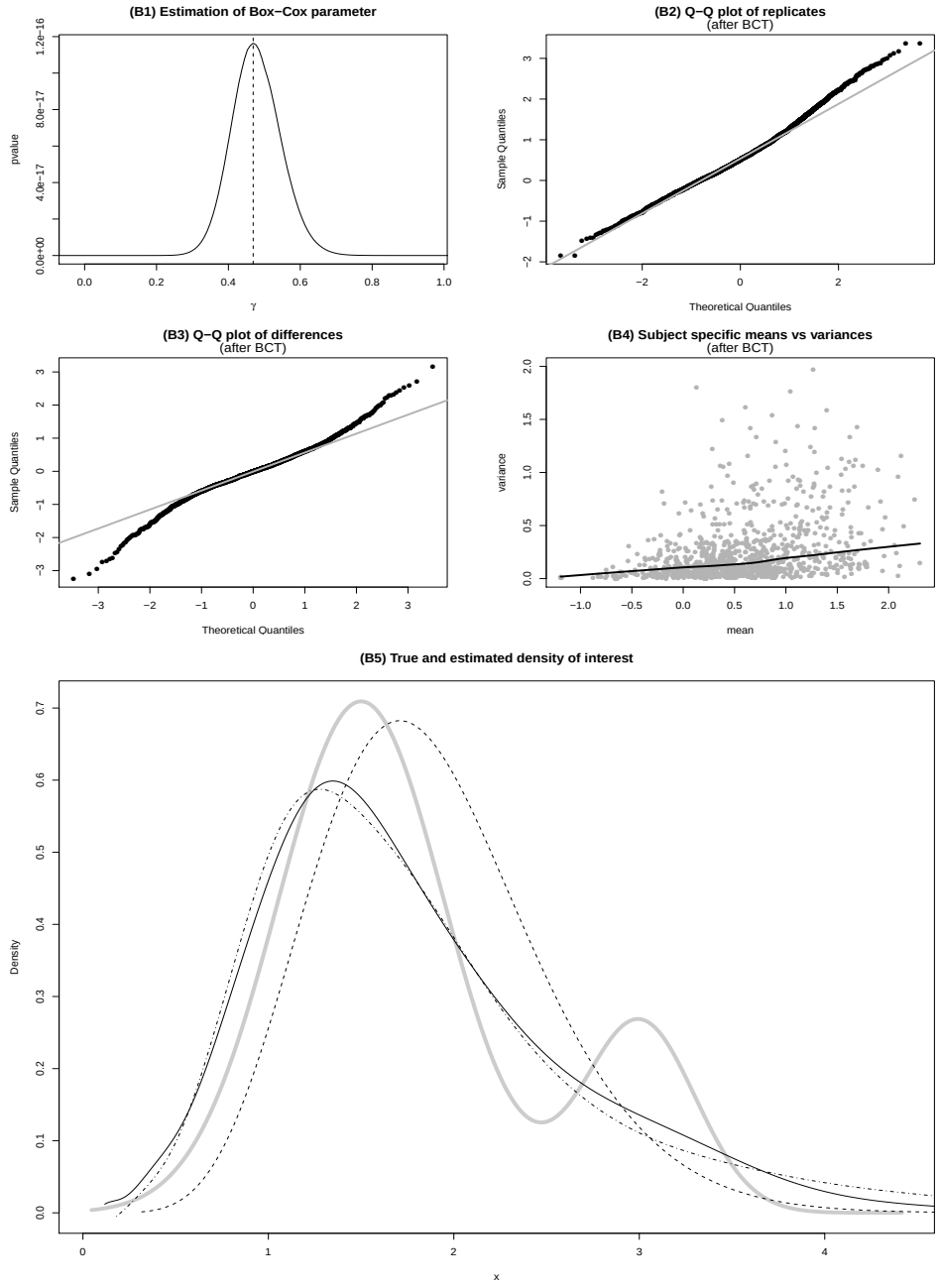

Figure A.3: Simulation results that illustrate the performance of the transform-retransform method that uses a Box-Cox transformation to make the differences normal by maximizing the p-value of Shapiro-Wilk test. Plot B1 shows the estimation of the Box-Cox transformation parameter; plot B2 shows the Q-Q plot of the transformed replicates; plot B3 shows the Q-Q plot of the differences of the transformed replicates; plot B4 shows the subject specific means and variances of transformed replicates; and plot B5 shows the estimated densities by the BIET method (dashed line), the DKET method (solid line) and the CHT method (dot-dashed lined) superimposed on the truth (bold gray line).

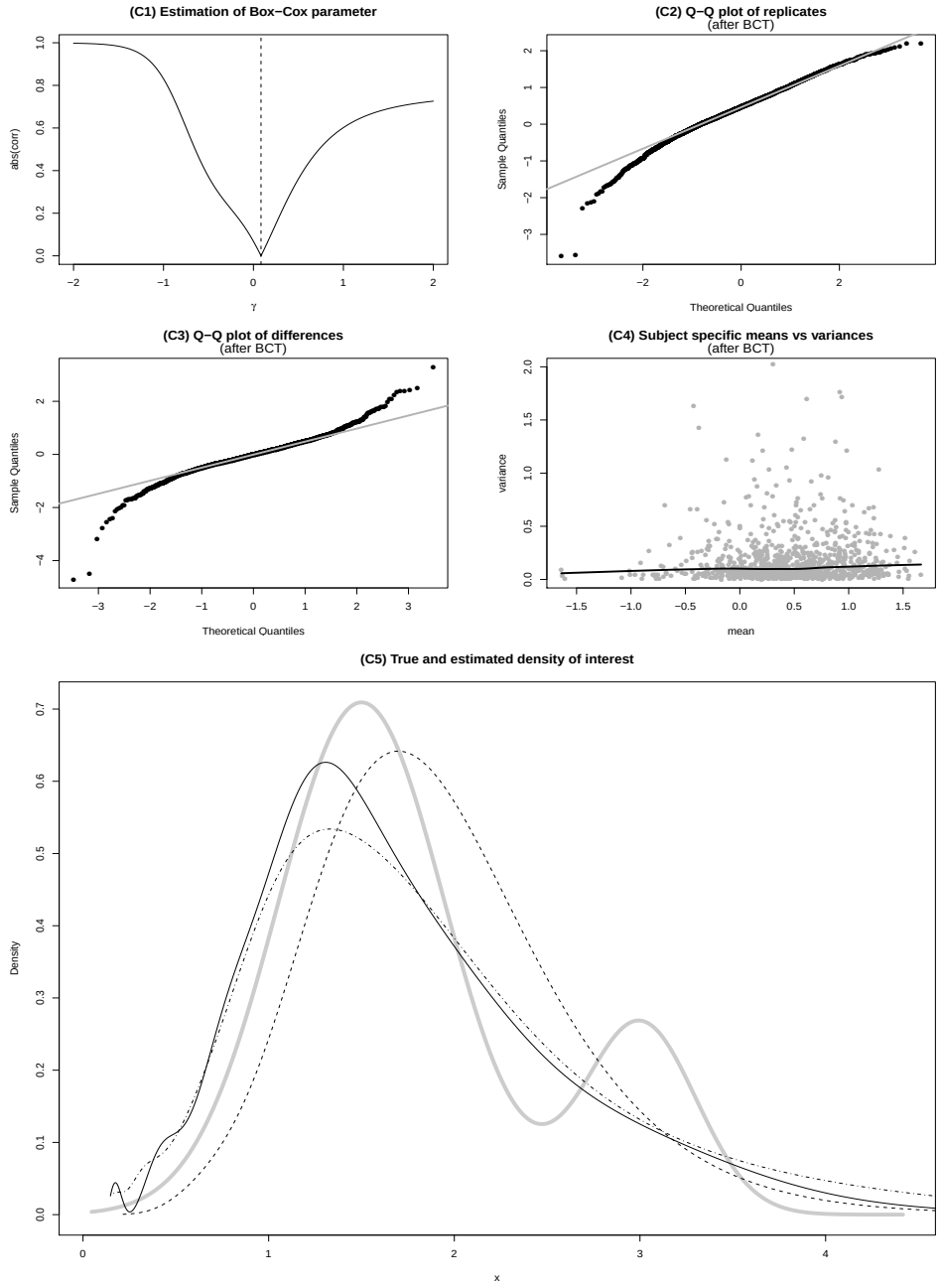

Figure A.4: Simulation results that illustrate the performance of the transform-retransform method that uses a Box-Cox transformation to make the errors independent of the variable of interest by minimizing the absolute value of the correlation between subject specific means and variances. Plot C1 shows the estimation of the Box-Cox transformation parameter; plot C2 shows the Q-Q plot of the transformed replicates; plot C3 shows the Q-Q plot of the differences of the transformed replicates; plot C4 shows the subject specific means and variances of transformed replicates; and plot C5 shows the estimated densities by the BIET method (dashed line), the DKET method (solid line) and the CHT method (dot-dashed lined) superimposed on the truth (bold gray line).

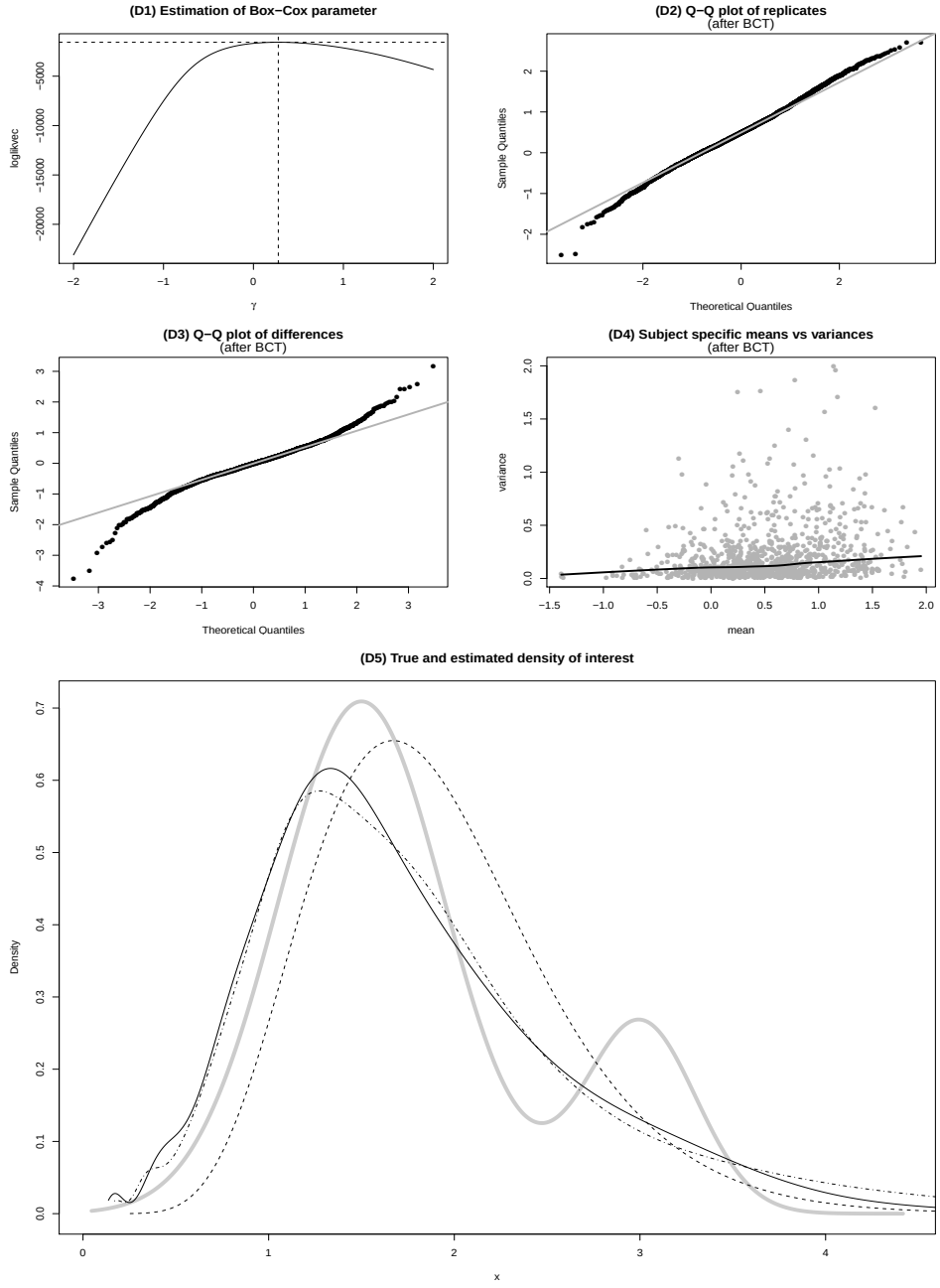

Figure A.5: Simulation results that illustrate the performance of the transform-retransform method that uses a Box-Cox transformation tries to make the observed replicates normal. Plot D1 shows the estimation of the Box-Cox transformation parameter; plot D2 shows the Q-Q plot of the transformed replicates; plot D3 shows the Q-Q plot of the differences of the transformed replicates; plot D4 shows the subject specific means and variances of transformed replicates; and plot D5 shows the estimated densities by the BIET method (dashed line), the DKET method (solid line) and the CHT method (dot-dashed lined) superimposed on the truth (bold gray line).

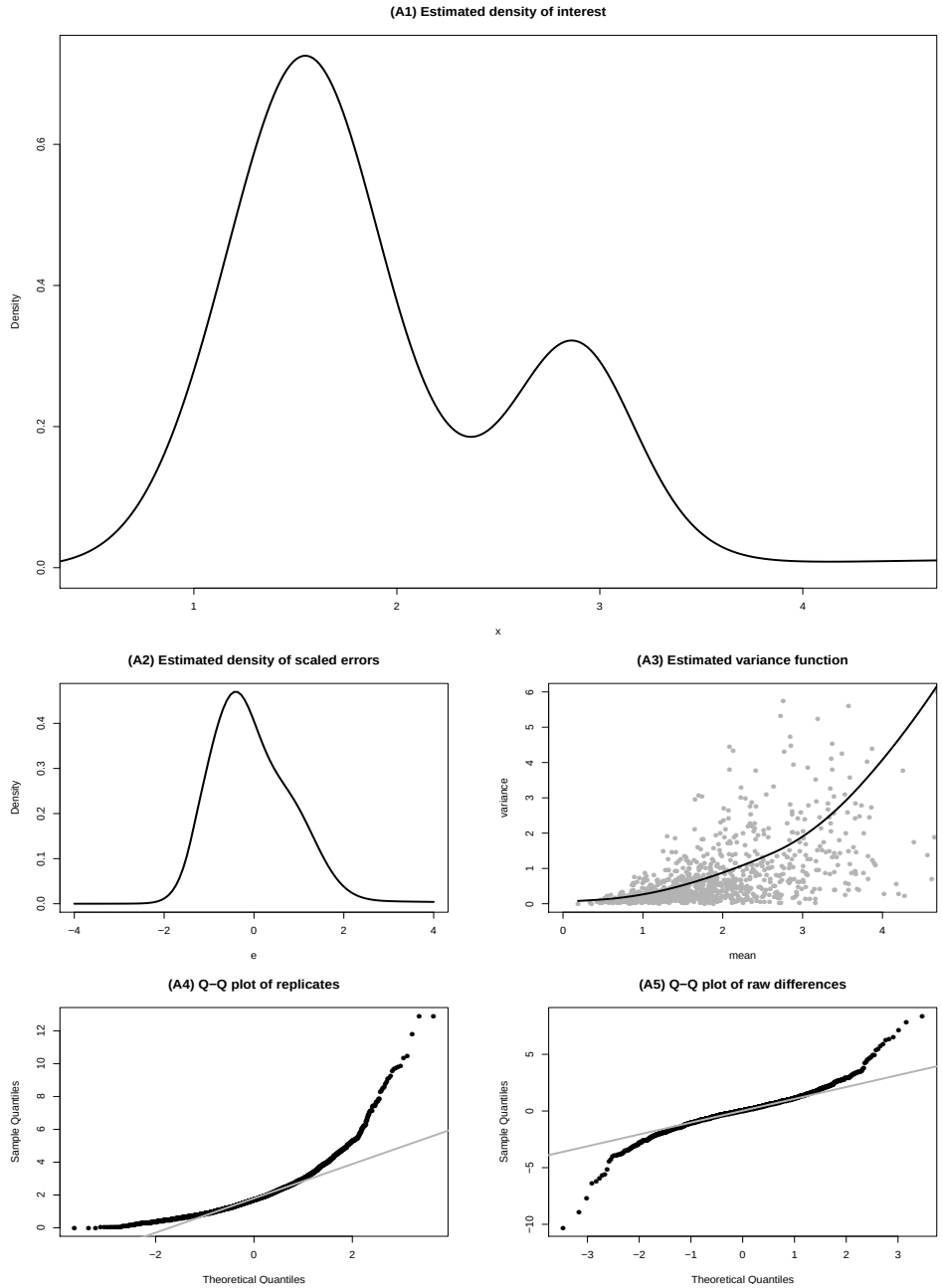

Figure A.6: Results produced by Model-III for daily intake of vitamin B6 from the EATS data set. Plot A1 shows the estimated density of interest; plot A2 shows the estimated density of measurement errors; plot A3 shows the estimated variance function superimposed with subject specific means and variances; plot A4 shows the Q-Q plot of the replicates; plot A5 shows the Q-Q plot of the differences of the replicates.

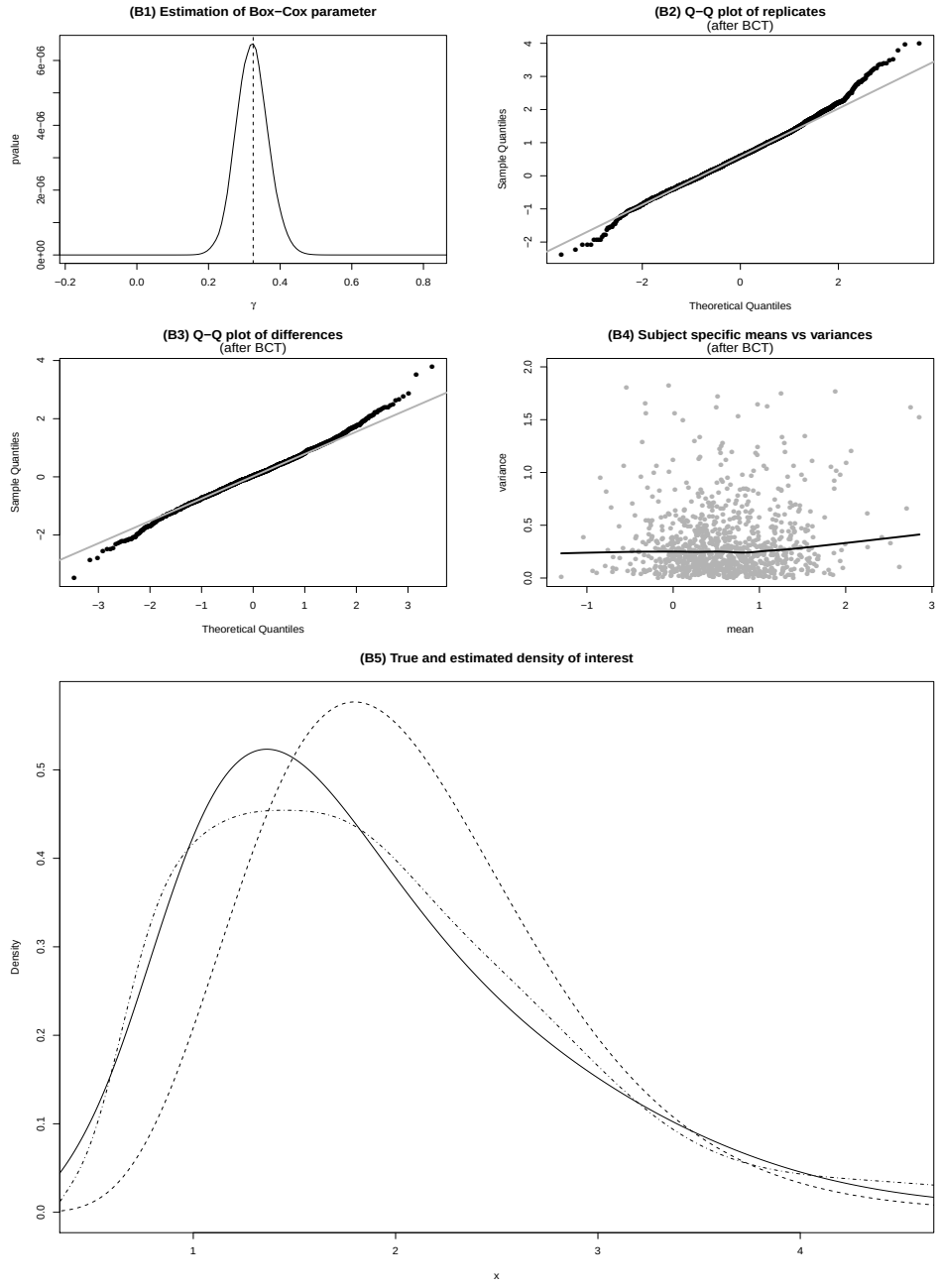

Figure A.7: Results for daily intake of vitamin B6 from the EATS data set produced by the transform-retransform method that uses Box-Cox transformation to make the differences normal by maximizing the p-value of Shapiro-Wilk test. Plot B1 shows the estimation of the Box-Cox transformation parameter; plot B2 shows the Q-Q plot of the transformed replicates; plot B3 shows the Q-Q plot of the differences of the transformed replicates; plot B4 shows the subject specific means and variances of transformed replicates; and plot B5 shows the estimated densities by the BIET method (dashed line), the DKET method (solid line) and the CHT method (dot-dashed lined).

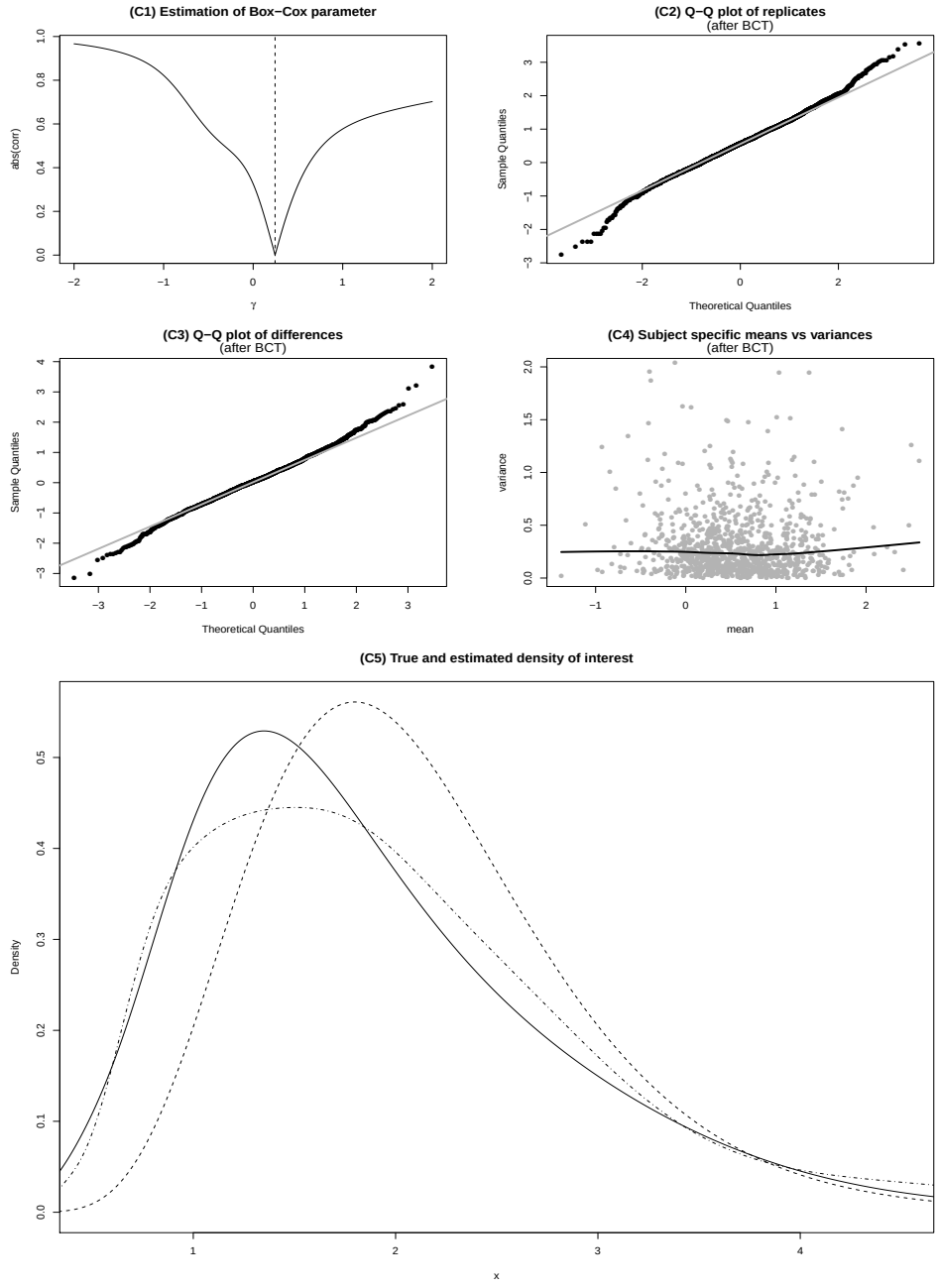

Figure A.8: Results for daily intake of vitamin B6 from the EATS data set produced by the transform-retransform method that uses Box-Cox transformation to make the errors independent of the variable of interest by minimizing the absolute value of the correlation between subject specific means and variances. Plot C1 shows the estimation of the Box-Cox transformation parameter; plot C2 shows the Q-Q plot of the transformed replicates; plot C3 shows the Q-Q plot of the differences of the transformed replicates; plot C4 shows the subject specific means and variances of transformed replicates; and plot C5 shows the estimated densities by the BIET method (dashed line), the DKET method (solid line) and the CHT method (dot-dashed lined).

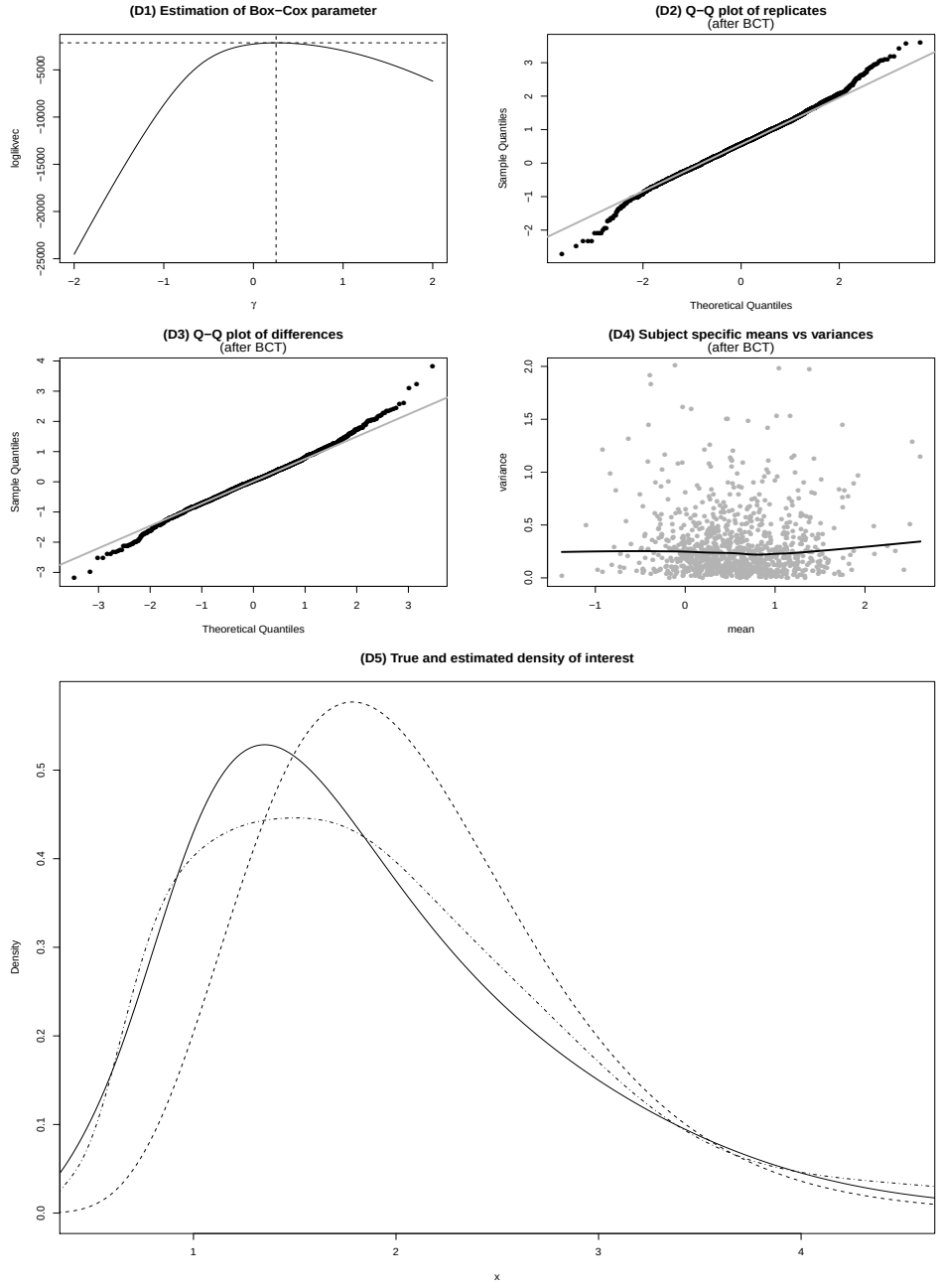

Figure A.9: Results for daily intake of vitamin B6 from the EATS data set produced by the transform-retransform method that uses Box-Cox transformation to make the observed replicates normal. Plot D1 shows the estimation of the Box-Cox transformation parameter; plot D2 shows the Q-Q plot of the transformed replicates; plot D3 shows the Q-Q plot of the differences of the transformed replicates; plot D4 shows the subject specific means and variances of transformed replicates; and plot D5 shows the estimated densities by the BIET method (dashed line), the DKET method (solid line) and the CHT method (dot-dashed lined).

APPENDIX B

APPENDIX TO SECTION 3

In this Appendix, we discuss the choice of hyper parameters and the details of the MCMC algorithm we used to draw samples from the posterior for the models described in 3. Figures B.1 and B.2 included in this Appendix show the posterior means associated with different percentiles of the estimates of MISE for two simulation instances and suggest that there is little simulation to simulation variability. To avoid unnecessary repetition, as in the main text, symbols sans the subscripts R, Y and W are used as generics for similar components and parameters of the models. For example, σ^2_ξ is a generic for $\sigma^2_{R,\xi}$, $\sigma^2_{Y,\xi}$ and $\sigma^2_{W,\xi}$; α_ϵ is a generic for α_{ϵ_Y} and α_{ϵ_W}; and so on.

B.1 Choice of Hyper-parameters

As in Section 2, we choose $\gamma_0 = 3$, $\nu_0 = 1/5$, $\mu_0 = \overline{\mathbf{W}}$, $\sigma_0^2 = S^2_{\mathbf{W}}(\gamma_0 - 1)/(1 + 1/\nu_0)$, $\sigma_{\widetilde{\mu}} = 3$, $a_\epsilon = 1$ and $b_\epsilon = 1$, $[A, B] = [\min(\overline{\mathbf{W}}_{1:n}) - 0.1 \text{ range}(\overline{\mathbf{W}}_{1:n}), \max(\overline{\mathbf{W}}_{1:n}) + 0.1 \text{ range}(\overline{\mathbf{W}}_{1:n})]$, $\alpha_X = 0.1$, $\alpha_\epsilon = 1$; and for modeling the variance functions $\widetilde{v}$ and the regression function r, quadratic (q=2) B-splines based are used. We set $\sigma^2_{R,\xi} = 0.1$ (also for the BCR method) and $\sigma^2_{Y,\xi} = \sigma^2_{W,\xi} = 0.01$. The smoothing parameters do not have a natural scale. Our experience with simulation studies suggests that, provided the true regression and variance functions are smooth, the data ranges are within 5 to 10 units and $15 \leq K_R \leq 25$ and $5 \leq K \leq 10$, the results are insensitive to our choice of the smoothing parameters and the number of knot points used. These observations are also in agreement with Ruppert (2002) who showed that while estimating a smooth function by penalized mixture of splines, after a minimum number of knots is used, further increases in the number of knots often have little effect on the fit. As in Section 2, in the simulation studies and the empirical application presented here, the number of knots are, therefore, kept fixed. In our implementation the B-splines for the regression function are based on $(2 \times 2 + 10 + 1) = 15$ knot points that divide the interval $[A, B]$ into 10 subintervals of equal length. The B-splines for the variance functions are based on $(2 \times 2 + 5 + 1) = 10$ knot points that divide the interval $[A, B]$ into 5 subintervals of equal length. For real data applications, we recommend scale transformations on the observed W and

Y values so that the range of these values be approximately 5 to 10 units.

B.2 Posterior Computation

As in Section 2.5, we define cluster labels $\mathbf{C}_{1:n}$, where $C_i = k$ if X_i is associated with the k^{th} component of f_X, modeled by (3.3). Also, define cluster labels $\{Z_{Y,i}\}_{i=1}^{n}$, where $Z_{Y,i} = k$ if $\epsilon_{Y,i}$ comes from the k^{th} component of f_{ϵ_Y}, modeled by (3.6). Similarly, define cluster labels $\{Z_{W,ij}\}_{i,j=1}^{n,m_i}$, where $Z_{W,ij} = k$ if $\epsilon_{W,ij}$ comes from the k^{th} component of $f_{W\epsilon}$, also modeled by (3.6). Given $Z_{Y,i} = k$,

$$\begin{aligned} f_{Y|X}(Y_{ij} \mid X_i, \boldsymbol{\xi}_R, \boldsymbol{\xi}_Y, p_{Y,k}, \mu_{Y,k1}, \mu_{Y,k2}, \sigma^2_{Y,k1}, \sigma^2_{Y,k2}) = \\ p_{Y,k}\text{Normal}\{Y_i \mid r(X_i, \boldsymbol{\xi}_R) + \widetilde{v}_Y(X_i, \boldsymbol{\xi}_Y)^{1/2}\mu_{Y,k1}, \widetilde{v}_Y(X_i, \boldsymbol{\xi}_Y)\sigma^2_{Y,k1}\} + \\ (1 - p_{Y,k})\text{Normal}\{Y_{ij} \mid r(X_i, \boldsymbol{\xi}_R) + \widetilde{v}_Y(X_i, \boldsymbol{\xi}_Y)^{1/2}\mu_{Y,k2}, \widetilde{v}_Y(X_i, \boldsymbol{\xi}_Y)\sigma^2_{Y,k2}\}. \end{aligned}$$

Given $Z_{W,ij} = k$, $f_{W|X}$ is similarly obtained. Let $N = \sum_{i=1}^{n} m_i$. Also, let $\boldsymbol{\theta}_{Y,k} = (p_{Y,k}, \mu_{Y,k1}, \mu_{Y,k2}, \sigma^2_{Y,k1}, \sigma^2_{Y,k2})^{\mathrm{T}}$ and $\boldsymbol{\theta}_{W,k} = (p_{W,k}, \mu_{W,k1}, \mu_{W,k2}, \sigma^2_{W,k1}, \sigma^2_{W,k2})^{\mathrm{T}}$. With a slight abuse of notation, define $\mathbf{W}_{1:N} = \{W_{ij}\}_{i,j=1}^{n,m_i}$ and $\mathbf{Z}_{W,1:N} = \{Z_{W,ij}\}_{i,j=1}^{n,m_i}$. In what follows, $\boldsymbol{\zeta}$ denotes a generic variable that collects all other parameters of a model, including $\mathbf{X}_{1:n}$, when not explicitly mentioned.

We tried two types of algorithms to fit the three DPMM components of our model - one exact method that integrates out the random mixture probabilities from the prior and posterior full conditionals of the cluster labels (Neal, 2000) as in Section 2.5 but with additional Metropolis-Hastings steps to draw samples from the full conditionals of $\boldsymbol{\xi}_R$ and $\boldsymbol{\xi}_Y$, and one that uses a weak limit approximation of the stick-breaking priors (Ishwaran and Zarepour, 2002) by finite dimensional symmetric Dirichlet priors $\boldsymbol{\pi} \sim \text{Dir}(\alpha/K, \ldots, \alpha/K)$, where K denotes the truncation level. There were no practical difference in the results produced by the two methods. We present here the latter that leads to simple Dirichlet full conditional for the mixture probabilities and multinomial full conditionals for the cluster labels. For most practical applications, a truncation level of 5 to 10 should suffice. We update other parameters of our model using the Metropolis-Hastings algorithm.

The starting values for $\mathbf{X}_{1:n}$ and $\boldsymbol{\xi}_W$ are determined exactly as described in Section 2.5. The C_i's are initialized at 1 with (μ_1, σ_1^2) set at the mean and the variance of the starting values. The Z's are all initialized at 1 with $(p_1, \widetilde{\mu}_1, \sigma^2_{11}, \sigma^2_{12}) = (0.5, 0, 1, 1)$. The initial value of $\boldsymbol{\xi}_R$ is obtained by fitting a naive regression model assuming the

regression errors $U_{Y,i}$ to be normally distributed with constant variance $\widehat{\sigma}_Y^2$. See B.3 for details. The MCMC iterations comprise the following steps.

1. **Updating the parameters of the distribution of X:** Conditionally given $\mathbf{X}_{1:n}$, the parameters specifying the DPMM for f_X can be updated using a Gibbs sampler. The full conditionals of $\boldsymbol{\pi}_X$ and C_i are given by

$$\begin{aligned} p(\boldsymbol{\pi}_X \mid \mathbf{C}_{1:n}, \boldsymbol{\zeta}) &\sim \text{Dir}(\alpha_X/K_X + n_{X,1}, \ldots, \alpha_X/K_X + n_{X,K_X}), \\ p(C_i = k \mid \mathbf{X}_{1:n}, \boldsymbol{\zeta}) &\propto \pi_{X,k}\ \text{Normal}(X_i \mid \mu_k, \sigma_k^2), \end{aligned}$$

   where $n_{X,k} = \sum_{i=1}^n 1_{\{C_i=k\}}$ is the number of C_i's that equal k.

   For all $k \in \mathbf{C}_{1:n}$, we update (μ_k, σ_k^2) using the closed-form joint full conditional given by $\{(\mu_k, \sigma_k^2) \mid \mathbf{X}_{1:n}, \boldsymbol{\zeta}\} = \text{NIG}(\mu_{nk}, \sigma_{nk}^2/\nu_{nk}, \gamma_{nk}, \sigma_{nk}^2)$, where $\nu_{nk} = (\nu_0 + n_{X,k})$; $\gamma_{nk} = (\gamma_0 + n_{X,k}/2)$; $\mu_{nk} = (\nu_0\mu_0 + n_{X,k}\sum_{\{i:C_i=k\}} X_i)/(\nu_0 + n_{X,k})$ and $\sigma_{nk}^2 = \sigma_0^2 + (\sum_{\{i:C_i=k\}} X_i^2 + \nu_0\mu_0^2 - \nu_{nk}\mu_{nk}^2)/2$. If $n_{X,k} = 0$, that is, if the k^{th} cluster is empty, we draw (μ_k, σ_k^2) from their prior.

2. **Updating $\mathbf{X}_{1:n}$:** Because the X_i's are conditionally independent, the full conditional of X_i is given by $p(X_i \mid \mathbf{W}_{1:N}, \boldsymbol{\zeta}) \propto \widehat{f}_X(X_i \mid \boldsymbol{\zeta}) \times f_{Y|X}(Y_i \mid X_i, \boldsymbol{\zeta}) \times \prod_{j=1}^{m_i} f_{W|X}(W_{ij} \mid X_i, \boldsymbol{\zeta})$. We use Metropolis-Hastings sampler to update the X_i's with proposal $q(X_i \to X_{i,new}) = TN(X_{i,new} \mid X_i, \sigma_X^2, [A, B])$, where $\sigma_X =$ (the range of $\overline{\mathbf{W}}_{1:n}$)/6 and $\text{TN}(\cdot \mid m, s^2, [\ell, u])$ denotes a truncated normal distribution with location m and scale s restricted to the interval $[\ell, u]$.

3. **Updating the parameters of the distribution of scaled errors:** The full conditionals of $\boldsymbol{\pi}_{\epsilon_Y}$ and $Z_{Y,i}$ are given by

$$\begin{aligned} p(\boldsymbol{\pi}_{\epsilon_Y} \mid \mathbf{Z}_{Y,1:n}, \boldsymbol{\zeta}) &\sim \text{Dir}(\alpha_{\epsilon_Y}/K_{\epsilon_Y} + n_{Y,1}, \ldots, \alpha_{\epsilon_Y}/K_{\epsilon_Y} + n_{Y,K_{\epsilon_Y}}), \\ p(Z_{Y,i} = k \mid \mathbf{X}_{1:n}, \boldsymbol{\zeta}) &\propto \pi_{\epsilon_Y,k}\ f_{Y|X}(Y_i \mid \boldsymbol{\theta}_{Y,k}, \boldsymbol{\zeta}), \end{aligned}$$

   where, for each i, $n_{Y,k} = \sum_{\ell=1}^n 1_{\{Z_{Y,\ell}=k\}}$, the number of $Z_{Y,\ell}$'s that equal k. If $k \notin \mathbf{Z}_{Y,1:n}$, we draw $\boldsymbol{\theta}_{Y,k}$ from the prior $p_0(\boldsymbol{\theta}_Y)$. For all $k \in \mathbf{Z}_{Y,1:n}$, we use a Metropolis-Hastings step to update $\boldsymbol{\theta}_{Y,k}$. We propose a new value for $\boldsymbol{\theta}_{Y,k}$ with the proposal $q(\boldsymbol{\theta}_{Y,k} \to \boldsymbol{\theta}_{Y,k,new}) = p_0(\boldsymbol{\theta}_{Y,k,new})$. We update $\boldsymbol{\theta}_{Y,k}$ to the proposed

value $\boldsymbol{\theta}_{Y,k,new}$ with probability

$$\min\left\{1, \frac{\prod_{\{i:Z_{Y,i}=k\}} f_{Y|X}(Y_i \mid \boldsymbol{\theta}_{Y,k,new}, \boldsymbol{\zeta})}{\prod_{\{i:Z_{Y,i}=k\}} f_{Y|X}(Y_i \mid \boldsymbol{\theta}_{Y,k}, \boldsymbol{\zeta})}\right\}.$$

We use the same algorithm to update $\mathbf{Z}_{W,1:N}$ and the $\boldsymbol{\theta}_{W,k}$'s. The full conditionals of $\boldsymbol{\pi}_{\epsilon_W}$ and $Z_{W,ij}$ are given by

$$\begin{aligned} p(\boldsymbol{\pi}_{\epsilon_Y} \mid \mathbf{Z}_{Y,1:N}, \boldsymbol{\zeta}) &\sim \mathrm{Dir}(\alpha_{\epsilon_W}/K_{\epsilon_W} + N_{W,1}, \ldots, \alpha_{\epsilon_Y}/K_{\epsilon_Y} + N_{W,K_{\epsilon_W}}), \\ p(Z_{W,ij} = k \mid \mathbf{X}_{1:n}, \boldsymbol{\zeta}) &\propto \pi_{\epsilon_W,k}\ f_{W|X}(W_{ij} \mid \boldsymbol{\theta}_{W,k}, \boldsymbol{\zeta}), \end{aligned}$$

where, $N_{W,k} = \sum_{rs} 1_{\{Z_{W,rs}=k\}}$, the number of $Z_{W,rs}$'s that equal k. If $k \notin \mathbf{Z}_{W,1:N}$, we draw $\boldsymbol{\theta}_{W,k}$ from the prior $p_0(\boldsymbol{\theta}_W)$. For all $k \in \mathbf{Z}_{W,1:N}$, we propose a new value for $\boldsymbol{\theta}_{W,k}$ with the same proposal $q(\cdot)$ and update $\boldsymbol{\theta}_{W,k}$ to the proposed value $\boldsymbol{\theta}_{W,k,new}$ with probability

$$\min\left\{1, \frac{\prod_{\{ij:Z_{W,ij}=k\}} f_{W|X}(W_{ij} \mid \boldsymbol{\theta}_{W,k,new}, \boldsymbol{\zeta})}{\prod_{\{ij:Z_{W,ij}=k\}} f_{W|X}(W_{ij} \mid \boldsymbol{\theta}_{W,k}, \boldsymbol{\zeta})}\right\}.$$

4. **Updating the parameters of the regression function:** The full conditional for $\boldsymbol{\xi}_R$ is given by $p(\boldsymbol{\xi}_R \mid \mathbf{Y}_{1:n}, \boldsymbol{\zeta}) \propto p_0(\boldsymbol{\xi}_R) \times \prod_{i=1}^n f_{Y|X}(Y_i \mid \boldsymbol{\xi}_R, \boldsymbol{\zeta})$. We use Metropolis-Hastings sampler to update $\boldsymbol{\xi}_R$ with proposal $q(\boldsymbol{\xi}_R \to \boldsymbol{\xi}_{R,new}) = \mathrm{MVN}(\boldsymbol{\xi}_{R,new} \mid \boldsymbol{\xi}_R, \Sigma_{R,\xi})$.

5. **Updating the parameters of the variance functions:** The full conditional for $\boldsymbol{\xi}_Y$ is given by $p(\boldsymbol{\xi}_Y \mid \mathbf{Y}_{1:n}, \boldsymbol{\zeta}) \propto p_0(\boldsymbol{\xi}_Y) \times \prod_{i=1}^n f_{Y|X}(Y_i \mid \boldsymbol{\xi}_Y, \boldsymbol{\zeta})$. We use Metropolis-Hastings sampler to update $\boldsymbol{\xi}_Y$ with random walk proposal $q(\boldsymbol{\xi}_Y \to \boldsymbol{\xi}_{Y,new}) = \mathrm{MVN}(\boldsymbol{\xi}_{Y,new} \mid \boldsymbol{\xi}_Y, \Sigma_{Y,\xi})$. Similarly, $\boldsymbol{\xi}_W$ is updated by Metropolis-Hastings sampler with proposal $q(\boldsymbol{\xi}_W \to \boldsymbol{\xi}_{W,new}) = \mathrm{MVN}(\boldsymbol{\xi}_{W,new} \mid \boldsymbol{\xi}_W, \Sigma_{W,\xi})$.

The covariance matrix $\Sigma_{R,\xi}$ of the proposal distribution for $\boldsymbol{\xi}_R$ is detailed in B.3. The initial choice of $\boldsymbol{\xi}_Y$ and the covariance matrix $\Sigma_{Y,\xi}$ of the proposal distribution for $\boldsymbol{\xi}_Y$ are discussed in B.4. The covariance matrix $\Sigma_{W,\xi}$ of the proposal distribution for $\boldsymbol{\xi}_W$ is taken to be the inverse of the negative Hessian matrix of $l(\boldsymbol{\xi}_W \mid 0.1, \overline{\mathbf{W}}_{1:n})$ evaluated at the chosen initial value of $\boldsymbol{\xi}_W$ exactly as in Appendix A.2.

B.3 Initial Values and Proposals for $\boldsymbol{\xi}_R$

The conditional posterior log-likelihood of $\boldsymbol{\xi}_R$ when $U_{Y,i} \sim \text{Normal}\{0, v_Y(X_i, \boldsymbol{\xi}_Y)\}$ is given by

$$\begin{aligned}
\ell(\boldsymbol{\xi}_R \mid \sigma^2_{R,\xi}, \boldsymbol{\xi}_Y, \mathbf{X}_{1:n}\} &= -\frac{1}{2\sigma^2_{R,\xi}}\boldsymbol{\xi}_R^{\mathrm{T}} P_R \boldsymbol{\xi}_R \\
&\quad -\frac{1}{2}\{\mathbf{Y}_{1:n} - \mathbf{B}_{q,J_R}(\mathbf{X}_{1:n})\boldsymbol{\xi}_R\}^{\mathrm{T}} V_Y^{-1}(\mathbf{X}_{1:n}, \boldsymbol{\xi}_Y)\{\mathbf{Y}_{1:n} - \mathbf{B}_{q,J_R}(\mathbf{X}_{1:n})\boldsymbol{\xi}_R\} \\
= \quad -\frac{1}{2}\boldsymbol{\xi}_R^{\mathrm{T}}\{\sigma^{-2}_{R,\xi} P_R + \mathbf{B}_{q,J_R}(\mathbf{X}_{1:n})^{\mathrm{T}} V_Y^{-1}(\mathbf{X}_{1:n}, \boldsymbol{\xi}_Y)\mathbf{B}_{q,J_R}(\mathbf{X}_{1:n})\}\boldsymbol{\xi}_R \\
&\quad -\boldsymbol{\xi}_R^{\mathrm{T}}\mathbf{B}_{q,J_R}(\mathbf{X}_{1:n})^{\mathrm{T}} V_Y^{-1}(\mathbf{X}_{1:n}, \boldsymbol{\xi}_Y)\mathbf{Y}_{1:n} - \frac{1}{2}\mathbf{Y}_{1:n}^{\mathrm{T}} V_Y^{-1}(\mathbf{X}_{1:n}, \boldsymbol{\xi}_Y)\mathbf{Y}_{1:n},
\end{aligned}$$

where $V_Y(\mathbf{X}_{1:n}, \boldsymbol{\xi}_Y) = \text{diag}\{v_Y(X_1, \boldsymbol{\xi}_Y), \ldots, v_Y(X_n, \boldsymbol{\xi}_Y)\}$. This implies

$$\{\boldsymbol{\xi}_R \mid \sigma^2_{R,\xi}, \boldsymbol{\xi}_Y, \mathbf{X}_{1:n}\} \sim \text{MVN}\{\boldsymbol{\mu}_{\boldsymbol{\xi}_R}(\mathbf{X}_{1:n}, \boldsymbol{\xi}_Y), \Sigma_{\boldsymbol{\xi}_R}(\mathbf{X}_{1:n}, \boldsymbol{\xi}_Y)\},$$

where

$$\begin{aligned}
\Sigma^{-1}_{\boldsymbol{\xi}_R}(\mathbf{X}_{1:n}, \boldsymbol{\xi}_Y) &= \{\sigma^{-2}_{R,\xi} P_R + \mathbf{B}_{q,J_R}(\mathbf{X}_{1:n})^{\mathrm{T}} V_Y^{-1}(\mathbf{X}_{1:n}, \boldsymbol{\xi}_Y)\mathbf{B}_{q,J_R}(\mathbf{X}_{1:n})\}, \\
\boldsymbol{\mu}_{\boldsymbol{\xi}_R}(\mathbf{X}_{1:n}, \boldsymbol{\xi}_Y) &= \Sigma^{-1}_{\boldsymbol{\xi}_R}(\mathbf{X}_{1:n}, \boldsymbol{\xi}_Y)\mathbf{B}_{q,J_R}(\mathbf{X}_{1:n}) V_Y^{-1}(\mathbf{X}_{1:n}, \boldsymbol{\xi}_Y)^{\mathrm{T}}\mathbf{Y}_{1:n}.
\end{aligned}$$

Let $\boldsymbol{\xi}_R^{(\star)} = \arg\min \sum_{i=1}^n \{Y_i - \mathbf{B}_{q,J_R}(\overline{W}_i)\boldsymbol{\xi}_R\}^2$ and further define $\widehat{\sigma}^2_Y(\boldsymbol{\xi}_R^{\star}, \overline{\mathbf{W}}_{1:n}) = \sum_{i=1}^n \{Y_i - \mathbf{B}_{q,J_R}(\overline{W}_i)\boldsymbol{\xi}_R^{(\star)}\}^2/(n - J_R)$. Numerical optimization to calculate $\boldsymbol{\xi}_R^{(\star)}$ is performed using the optim routine in R. In the conditionally heteroscedastic case, $\widehat{\sigma}^2_Y$ may be taken to be a crude estimate of the 'average' variance of U_Y over possible values of X. The initial value for the MCMC iterations for $\boldsymbol{\xi}_R$ is then obtained as $\boldsymbol{\xi}_R^{(0)} = \mu_{\boldsymbol{\xi}_R}(\overline{\mathbf{W}}_{1:n}, \boldsymbol{\xi}_Y^{(0)})$, where $\xi^{(0)}_{Y,j} = \log \widehat{\sigma}^2_Y$ for all j. Let $\mathbf{X}_{1:n}^{(m)}$ and $\boldsymbol{\xi}_Y^{(m)}$ denote the sampled values of $\mathbf{X}_{1:n}$ and $\boldsymbol{\xi}_Y$, respectively, for the m^{th} MCMC iteration. Also let $\text{var}(\epsilon_Y^{(m)})$ the estimated variance of ϵ_Y for the m^{th} MCMC iteration. For a few initial iterations the covariance matrix of the proposal for $\boldsymbol{\xi}_R$ is taken to be $\Sigma^{(m)}_{R,\xi} = [\sigma^{-2}_{R,\xi} P_R + \mathbf{B}^{\mathrm{T}}_{q,J_R}(\mathbf{X}^{(m)}) V^{-1}_{Y(m)} \mathbf{B}_{q,J_R}(\mathbf{X}^{(m)})]^{-1}$, where $V_{Y(m)} = \text{diag}\{v_Y(X_1^{(m)}, \boldsymbol{\xi}_Y^{(m)})\text{var}(\epsilon_Y^{(m)}), \ldots, v_Y(X_n^{(m)}, \boldsymbol{\xi}_Y^{(m)})\text{var}(\epsilon_Y^{(m)})\}$. After that the $\Sigma^{(m)}_{R,\xi}$ becomes stable and is no longer updated.

B.4 Initial Values and Proposals for $\boldsymbol{\xi}_Y$

Obtaining a good initial idea about conditional heteroscedasticity in regression errors is difficult. Each $\xi_{Y,j}$ is thus initialized at $\xi_{Y,j}^{(0)} = \log \widehat{\sigma}_Y^2$.

We use Metropolis-Hastings sampler to update $\boldsymbol{\xi}_Y$ with random walk proposal $q(\boldsymbol{\xi}_Y \to \boldsymbol{\xi}_{Y,new}) = \text{MVN}(\boldsymbol{\xi}_{Y,new} \mid \boldsymbol{\xi}_Y, \Sigma_{Y,\xi})$. The covariance matrix $\Sigma_{Y,\xi}$ of the proposal distribution for $\boldsymbol{\xi}_Y$ is taken to be $\Sigma_{Y,\xi}(i,j) = \sigma_{\xi_Y}^2 \rho_{\xi_Y}^{|i-j|}$. For most practical applications the variance function v_Y may be assumed to be smooth. Hence, the components of $\boldsymbol{\xi}_Y$ may be expected to be highly correlated and the value of ρ_{ξ_Y} is chosen to be $\rho_{\xi_Y} = 0.9$. The parameter σ_{ξ_Y} is tuned to get good acceptance rates for the Metropolis-Hastings sampler.

B.5 Additional Simulation Results

B.5.1 Comparison with a Possible Nonparametric Alternative

The model for the densities of the regression errors and the measurement errors that we described assumes a multiplicative structural assumption (3.5). As in Section 2.6.2, we ran some simulations to study the performance of the model under violations of this assumption.

Suppressing the suffixes Y and W, the true conditional distributions that generate the regression errors and the measurement errors are designed to be of the generic form $f_{U|X}(U \mid X) = \sum_{k=1}^K \pi_k(X) f_{cU}(U \mid \sigma_k^2, \boldsymbol{\theta}_k)$, where each of the K component densities has mean zero, the k^{th} component has variance σ_k^2, and $\boldsymbol{\theta}_k$ includes additional parameters. Through the mixture probabilities $\pi_k(X)$, X affects all aspects of these mixture densities. The conditional ℓ^{th} order central moment of the mixture density is given by $\mu_\ell(X, \{\sigma_k^2, \boldsymbol{\theta}_k\}_{k=1}^K) = \sum_{k=1}^K \pi_k(X) \mu_{\ell,k}(\sigma_k^2, \boldsymbol{\theta}_k)$, where $\mu_{\ell,k}(\sigma_k^2, \boldsymbol{\theta}_k)$ is the ℓ^{th} order central moment of the k^{th} component density. In particular, the conditional variance is given by $\text{var}(U \mid X) = \sum_{k=1}^K \pi_k(X) \sigma_k^2$. The truth, therefore, departs from the multiplicative structural assumption (3.5).

The performance of our model is evaluated for two sample sizes $n = 500, 1000$; two choices for the number of surrogates per subject $m = 2, 3$; one density of the covariate $f_X(X) = 0.8\ \text{Normal}(X \mid -1, 0.5) + 0.2\ \text{Normal}(X \mid 1, 0.5)$; one regression function $r(X) = \sin(\pi X/2)/[1 + 2X^2\{\text{sign}(X) + 1\}]$; and three different choices for the component densities f_{cU} - (a) $\text{Normal}(0, \sigma_k^2)$, (b) $\text{SN}(\cdot \mid 0, \sigma_k^2, \lambda)$ and (c) $\text{SN}(\cdot \mid 0, \sigma_k^2, \lambda_k)$. Here $\text{SN}(\cdot \mid 0, \sigma^2, \lambda)$ denotes a skew normal density with mean zero, variance σ_k^2 and skewness λ (Azzalini, 1985). In each case, $K = 7$. The

component specific variance parameters of the density of the regression errors are set by minimizing the sum of squares $g_Y(X) = \{(0.3 + X/8)^2 - \sum_{k=1}^{K} \pi_{Y,k}(X)\sigma^2_{Y,k}\}$ on a grid. Similarly, the component specific variance parameters of the density of the measurement errors are set by minimizing the sum of squares $g_W(X) = \{(0.8 + X/4)^2 - \sum_{k=1}^{K} \pi_{W,k}(X)\sigma^2_{W,k}\}$ on a grid. For the case (b) we set $\lambda = 7$. For the case (c) λ_k's take values in $\{7, 3, 1, 0, -1, -3, -7\}$ with λ_k decreasing as X increases.

The estimated MISEs for these misspecified cases are presented in Table B.1. When the estimated MISEs are comparable to the MISEs for normally distributed regression and measurement errors reported in Table 3.4. These results suggest that our method is fairly robust to departures from the multiplicative structural assumption (3.5).

True Error Distribution	Sample Size	Number of Replicates	MISE ×100
(a)	500	2	6.83
		3	6.70
	1000	2	6.65
		3	5.91
(b)	500	2	11.43
		3	6.27
	1000	2	6.38
		3	5.70
(c)	500	2	6.60
		3	6.08
	1000	2	5.58
		3	4.26

Table B.1: Mean Integrated Square Error (MISE) performance of our model for simulation experiments when true error generating densities depart from the multiplicative structural assumption (3.5). See Section B.5 for details.

B.5.2 Comparison with an Improved Parametric Alternative

The simulation results presented in Section 3.3.1 showed that our more flexible method outperforms the parametric BCR method of Berry, et al. (2002), that as-

sumes homoscedasticity and normality of regression and measurement errors, even when these assumptions are satisfied. Berry, et al. (2002) used truncated polynomial splines or P-splines to model the regression function. We used B-splines to model the regression and the variance functions. As also mentioned in Section 3.3.1, B-splines are locally supported and nearly orthogonal and hence numerically more stable than P-splines. This increased numerical stability of our model results in better performance even in situations when the parametric assumption of the BCR model are satisfied. We present here the results of additional simulation experiments that provide support this claim. We compare our method with an improved version of the BCR method, referred to as the BCRB method, that makes the same parametric assumptions as the BCR model but uses B-splines, not P-splines, to model the regression function. We considered two subcases from each of the three scenarios considered in Section 3.3.

1. **Cases from Table 3.2:** X is normally distributed, measurement errors are normally distributed and homoscedastic, regression errors are homoscedastic.

   (**A**) **Case 2A:** Regression errors are normally distributed. All parametric assumptions of the BCR method are satisfied in this case.

   (**B**) **Case 2B:** Regression errors are distributed according to heavy-tailed symmetric distribution 1. (density (d) in Table 3.1 and Figure 3.1).

2. **Cases from Table 3.3:** X is not normally distributed, measurement errors are homoscedastic, regression errors are homoscedastic.

   (**A**) **Case 3A:** Regression and measurement errors are distributed according to a light-tailed bimodal skewed density (density (b) in Table 3.1 and Figure 3.1).

   (**B**) **Case 3B:** Regression and measurement errors are distributed according to a light-tailed bimodal symmetric density (density (c) in Table 3.1 and Figure 3.1).

3. **Cases from Table 3.4:** X is not normally distributed, measurement errors are conditionally heteroscedastic, regression errors are conditionally heteroscedastic.

(**A**) **Case 4A:** Regression errors are distributed according to heavy-tailed symmetric distribution 1. (density (d) in Table 3.1 and Figure 3.1).

(**B**) **Case 4B:** Regression errors are distributed according to heavy-tailed symmetric distribution 2. (density (e) in Table 3.1 and Figure 3.1).

Results are presented in Table B.2. As expected, when the parametric assumptions of the BCR model were true, the BCRB method outperformed our method. In all other cases, our method outperformed the BCRB method.

	Sample Size	Number of Replicates	MISE ×100				
			BCR	BCRB	BSP	Naive	DKE
Case 2A	500	2	4.98	**2.22**	2.84	16.66	24.85
		3	4.09	**1.67**	1.82	11.97	22.84
	1000	2	3.11	**1.29**	1.53	18.05	20.21
		3	2.42	**0.96**	0.96	10.88	16.64
Case 2B	500	2	4.78	2.41	**1.82**	16.66	24.85
		3	4.09	1.83	**1.38**	11.97	22.84
	1000	2	2.87	1.47	**1.10**	18.05	20.21
		3	2.38	1.21	**0.76**	10.88	16.64
Case 3A	500	2	19.52	12.01	**4.66**	34.44	46.67
		3	16.20	9.78	**2.84**	23.86	38.36
	1000	2	14.01	9.44	**2.61**	33.57	37.64
		3	11.79	7.34	**1.55**	23.22	33.30
Case 3B	500	2	20.18	12.82	**5.09**	34.67	45.97
		3	17.15	10.26	**3.20**	24.08	37.54
	1000	2	15.73	11.37	**2.67**	31.61	38.95
		3	13.01	9.44	**1.87**	22.52	32.56
Case 4A	500	2	26.44	13.39	**2.93**	12.84	65.74
		3	15.80	10.93	**2.07**	9.56	43.57
	1000	2	23.89	13.46	**1.49**	13.40	60.53
		3	15.42	13.35	**1.05**	10.05	39.29
Case 4B	500	2	28.58	19.56	**6.11**	16.38	51.92
		3	20.01	14.99	**3.89**	11.65	40.99
	1000	2	26.73	16.72	**3.44**	15.16	47.53
		3	18.57	13.99	**2.31**	10.45	35.83

Table B.2: Mean Integrated Square Error (MISE) performance of our model (BSP) compared to the BCRB (the model of Berry, et al., 2002 but with B-splines) for five different scenarios. Two sub cases from each of Table 3.2, Table 3.3 and Table 3.4 in the main text are considered. Results produced by the BCR method, the naive method and the DKE method are also shown. See Section 3.3 of the main text for additional details.

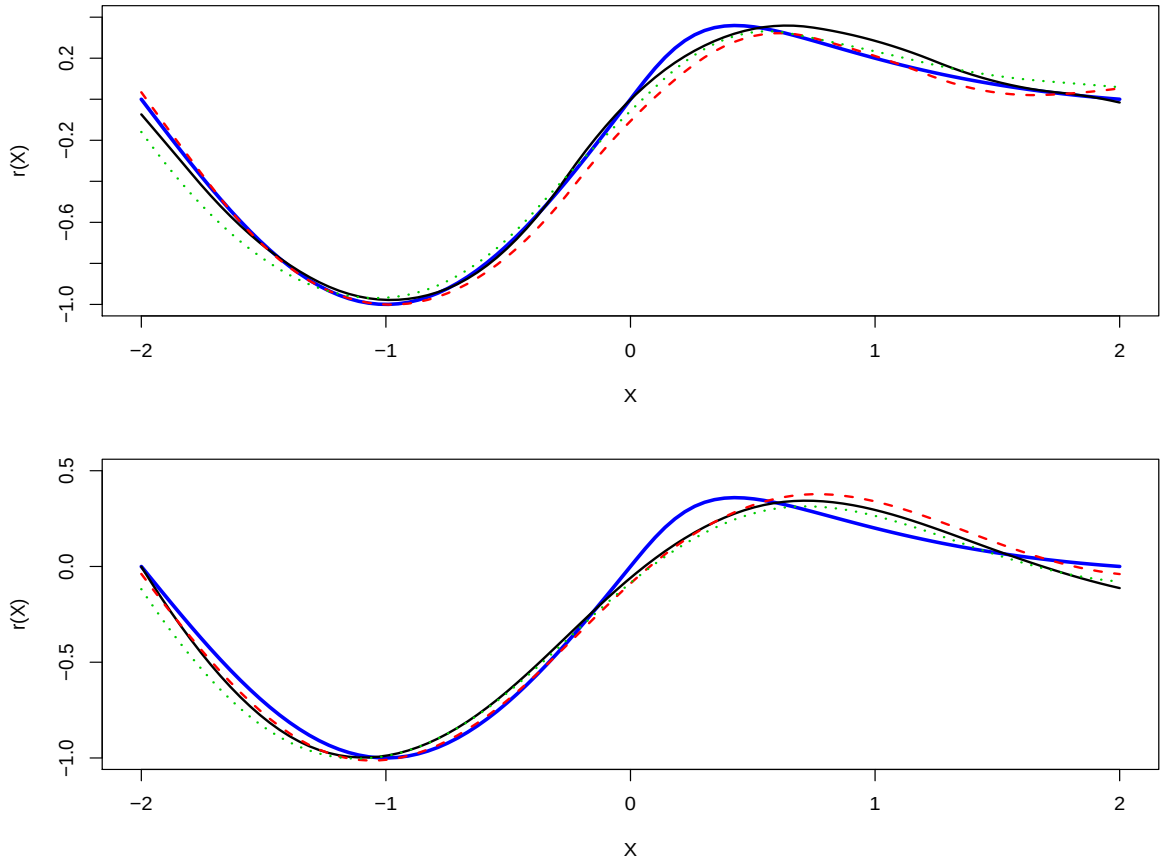

Figure B.1: Plot showing the truth and the posterior means associated with different percentiles of estimates of MISE for our method (BSP, top panel) and the method of Berry, Carroll and Ruppert (2002) (BCR, bottom panel) methods when the covariate $X \sim \text{Normal}(0, 1)$, the regression function is given by $r(X) = \sin(\pi X/2)/[1 + 2X^2\{\text{sign}(X) + 1\}]$, the regression errors $U_Y \sim \text{Normal}(0, 0.3^2)$, and the measurement errors $U_W = \text{Normal}(0, 0.8)^2$, sample size $n = 1000$ and $m = 3$ replicates per subject. The solid blue curves show the truth, the black solid curves show the posterior means that correspond to the 25^{th} percentile, the red dashed curves show the posterior means that correspond to the 50^{th} percentile, and the green dotted curves show the posterior means that correspond to the 75^{th} percentile.

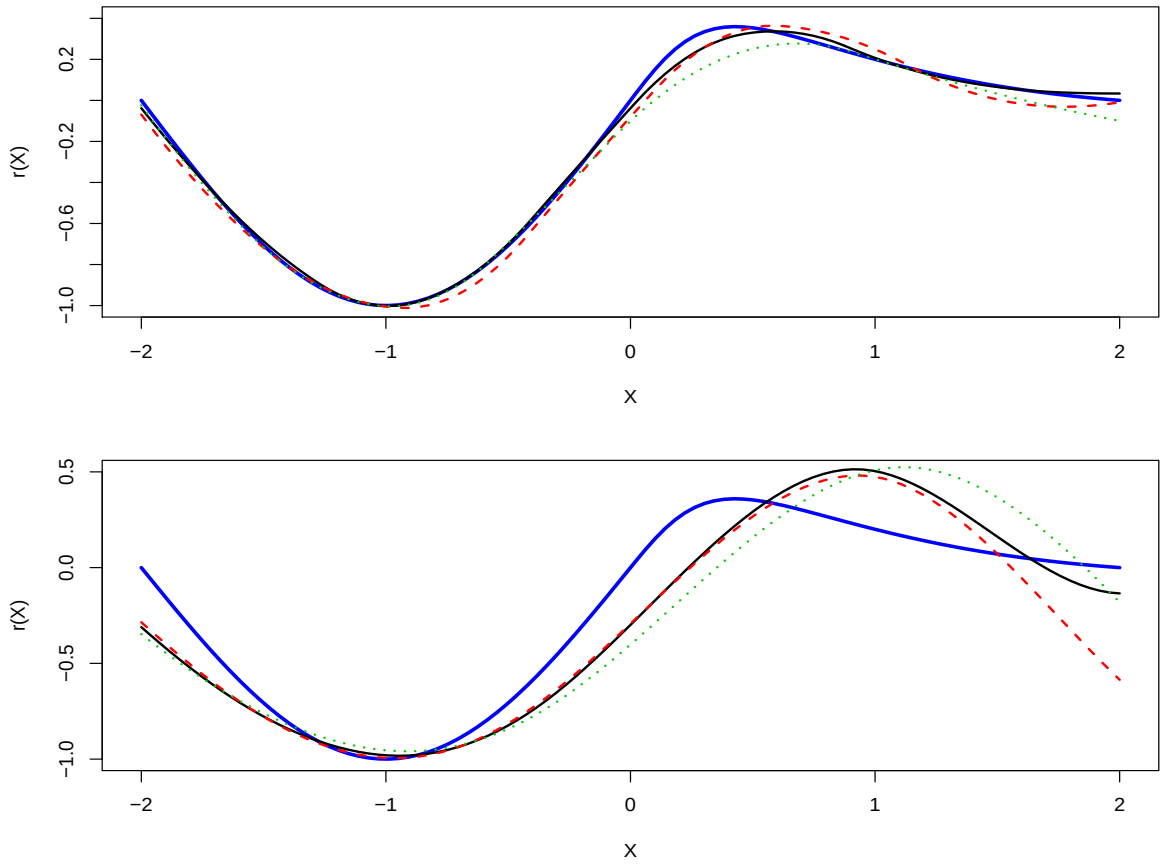

Figure B.2: Plot showing the truth and the posterior means associated with different percentiles of estimates of MISE for our method (BSP, top panel) and the method of Berry, Carroll and Ruppert (2002) (BCR, bottom panel) methods when the covariate $X \sim 0.8$ Normal$(-1, 0.5) + 0.2$ Normal$(1, 0.5)$, $X \sim$ Normal$(0, 1)$, the regression function is given by $r(X) = \sin(\pi X/2)/[1 + 2X^2\{\text{sign}(X) + 1\}]$, the regression errors $U_Y \sim v_Y^{1/2}(X)\epsilon_Y$ with $v_Y(X) = (0.3 + X/8)^2$, and the measurement errors $U_W = v_W^{1/2}(X)\epsilon_W$ with $v_W(X) = (0.8 + X/4)^2$, ϵ_Y and ϵ_W both follow the heavy tailed error density (d) depicted in Figure 3.1, sample size $n = 1000$ and $m = 3$ replicates per subject. The solid blue curves show the truth, the black solid curves show the posterior means that correspond to the 25^{th} percentile, the red dashed curves show the posterior means that correspond to the 50^{th} percentile, and the green dotted curves show the posterior means that correspond to the 75^{th} percentile.

APPENDIX C

APPENDIX TO SECTION 4

C.1 Finite Mixture Models vs Infinite Mixture Models

In Section 4, we modeled the density of interest $f_{\mathbf{X}}$ and the density of the scaled measurement errors $f_{\boldsymbol{\epsilon}}$ using mixtures of fixed finite number of multivariate normal kernels. Alternative approaches that escape the need to prespecify the number of mixture components include models with potentially infinite number of mixture components, models induced by Dirichlet processes (Ferguson, 1973; Escobar and West, 1995) being perhaps the most popular among such techniques. Apart from flexibility, one major advantage of such techniques comes from the ability of associated MCMC machinery to perform model selection and model averaging implicitly and semiautomatically. Model averaging is achieved by allowing the number of mixture components to vary from one MCMC iteration to the other. The number of mixture components that is visited the maximum number of times by the sampler then provides a maximum a-posteriori (MAP) estimate of the number of mixture components required to approximate the target density. However, in complicated multivariate set up like ours, MCMC algorithms for such infinite dimensional models become computationally highly intensive. Mixtures based on fixed finite number of components, on the other hand, can greatly reduce computational complexity. Recent studies of asymptotic properties of the posterior of overfitted mixture models (Rousseau and Mengersen, 2011) suggest that mixture models with sufficiently large number of components can perform automatic model selection and model averaging just like infinite dimensional models. Additionally, as the proofs of the results in Section D.3 imply, the use of mixture models with fixed finite number of components does not necessarily imply a compromise on the issue of flexibility. The approaches adopted in this dissertation try to take the best from both worlds. Computational burden is reduced by keeping the number of mixture components fixed at some finite values. At the same time, simultaneous semiautomatic model selection and model averaging is achieved by exploiting asymptotic properties of overfitted mixture models. We elaborate our arguments below, pointing out the close connections and the subtle differences our adopted finite dimensional models have with the aforementioned

infinite dimensional alternatives.

C.1.1 Computational Complexity

The finite mixtures of multivariate normal kernels with symmetric Dirichlet priors that we used in Section 4 to model both the density of interest $f_{\mathbf{X}}$ and the density of the scaled measurement errors $f_{\boldsymbol{\epsilon}}$ have close connections with infinite dimensional Dirichlet process based mixture models. Indeed, as $K_{\mathbf{X}} \to \infty$, the finite dimensional symmetric Dirichlet prior (4.3) used to model $f_{\mathbf{X}}$ converges weakly to a countably infinite distribution distributed according to a Dirichlet process with concentration parameter $\alpha_{\mathbf{X}}$ (Ishwaran and Zarepour, 2000, 2002). Our proposed mechanism to enforce the mean zero restriction on $f_{\boldsymbol{\epsilon}}$ specifically requires a finite dimensional symmetric prior on the mixture probabilities and therefore does not admit a straightforward infinite dimensional extension. But in the limit, as $K_{\boldsymbol{\epsilon}} \to \infty$, a reformulation of the model results in a complicated multivariate version of the infinite dimensional model of Sarkar, et al. (2014) (See Lemma 1 in Section 4). The implementation of such complex models, specially the complicated mean restricted model for the scaled errors, will be computationally intensive in a multivariate setting like ours. The computational simplicity of the finite dimensional methods proposed in Section 4 make them particularly suitable for multivariate problems.

In this paragraph, we discuss additional mixing issues that render infinite dimensional models, particularly the ones with non or semiconjugate priors on the component specific parameters (like our MLFA model), unsuitable for multivariate applications. There are two main types of MCMC algorithms for fitting infinite dimensional mixture models - conditional methods and marginal methods. In the conditional scheme, the mixture probabilities are sampled. The mixture labels are then updated independently, conditional on the mixture probabilities. The mixture probabilities in infinite dimensional mixture models can be stochastically ordered. For instance, mixture probabilities in a Dirichlet process mixture model satisfy $E(\pi_k) > E(\pi_{k+1})$ and $\Pr(\pi_k > \pi_{k+1}) > 0.5$ for all $k \in \mathbb{N}$. This imposes weak identifiability on the mixture labels resulting in a complicated model space comprising many local modes of varying importance. Different permutations of the mixture labels are not equivalent and exploration of the entire model space becomes important for valid inference. In high dimensional and large data settings this is difficult to achieve even by sophisticated MCMC algorithms with carefully designed label switching moves (Hastie, et

al., 2013). The problem can be avoided with marginal methods (Neal, 2000) that integrate out the mixture probabilities and work with the resulting Polya urn scheme, rendering the mixture labels dependent but nonidentifiable. Unfortunately, such integration is possible only when conjugate priors are assigned to the component specific parameters. Typically for infinite dimensional models with non or semiconjugate priors on the component specific parameters, good mixing is thus difficult to achieve, particularly in complicated multivariate setup like ours. On the contrary, the issues of mixing and convergence become much less important for finite mixture models with exchangeable priors on the mixture probabilities. With $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$ mixture components for the densities $f_{\mathbf{X}}$ and $f_{\boldsymbol{\epsilon}}$, respectively, the posterior is still multimodal but comprises $K_{\mathbf{X}}! \times K_{\boldsymbol{\epsilon}}!$ modal regions that are exact copies of each other. For inference on the overall density or any other functions of interest that are invariant to permutations of the mixture labels, it is only important that the MCMC sampler visits and explores at least one of the modal regions well and label switching (or the lack of it) does not present any problem (Geweke, 2007).

C.1.2 Model Selection and Model Averaging

As mentioned at the beginning of Section C.1, a major advantage of infinite dimensional mixture models is their ability to implicitly and semiautomatically perform model selection and model averaging. Properties of overfitted mixture models can be exploited to achieve the same in finite dimensional models with sufficiently large number of components. Recently Rousseau and Mengersen (2011) studied the asymptotic behavior of the posterior for overfitted mixture models with Dirichlet prior $\text{Dir}(\alpha_1, \ldots, \alpha_K)$ on the mixture probabilities in a measurement error free set up and showed that the hyper parameter $(\alpha_1, \ldots, \alpha_k)$ strongly influences the way the posterior handles overfitting. In particular, when $\max_{k=1,\ldots,K} \alpha_k < L/2$, where L denotes the number of parameters specifying the component kernels, the posterior is asymptotically stable and concentrates in regions with empty redundant components. In Section 4, we chose symmetric Dirichlet priors $\text{Dir}(\alpha/K, \ldots, \alpha/K)$ on the mixture probabilities to model both the density of interest $f_{\mathbf{X}}$ and the density of the scaled measurement errors $f_{\boldsymbol{\epsilon}}$. We set $\alpha_{\mathbf{X}} = \alpha_{\boldsymbol{\epsilon}} = 1$ so that the condition $\alpha/K < L/2$ is satisfied for both $f_{\mathbf{X}}$ and $f_{\boldsymbol{\epsilon}}$. In simulation experiments reported in Section 4.8, the behavior of the posterior was similar to that observed by Rousseau and Mengersen (2011) in measurement error free set up. That is, when $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$ were assigned

sufficiently large values, the MCMC chain quickly reached a stable stage where the redundant components became empty. See Figure 4.7 and Figure C.6 and C.7 for illustrations. Since such overfitted mixture models allow the number of nonempty mixture components to vary from one MCMC iteration to the next, model averaging is automatically achieved. MAP estimates of the numbers of mixture components required to approximate the target densities are given by the numbers of components which are visited the maximum number of times by the MCMC sampler, as in the case of infinite mixture models.

As discussed in the main text, for the MIW method, when the measurement errors are conditionally heteroscedastic and the true covariance matrices are highly sparse, the strategy usually overestimates the number of non-empty mixture components required to approximate the target densities well. In these cases, the MIW method becomes highly numerically unstable and much larger sample sizes are required for the asymptotic results to hold. See Figure 4.6 for an illustration. This may be regarded more as a limitation of the MIW method than a limitation of the adopted strategy to determine $K_{\mathbf{X}}$ and $K_{\boldsymbol{\epsilon}}$. For the numerically more stable MLFA model, the asymptotic results hold for much smaller sample sizes and such models are also more robust to overestimation of the number of nonempty clusters.

C.1.3 Model Flexibility

The proofs of the support results presented in Section D.3 require that the number of mixture components of the corresponding mixture models be allowed to vary over the set of all positive integers. However, as the technical details of the proofs reveal, the use of mixture models with fixed finite number of components does not necessarily imply a compromise on the issue of flexibility. Indeed, a common recurring idea in the proofs of all these results, including those for the variance functions, is to show that any function coming from the target class can be approximated with any desired level of accuracy by the corresponding finite mixture models provided the models comprise sufficiently large number of mixture components and the function satisfies some fairly minimal regularly conditions. The requirement that the priors on the number of mixture components assign positive probability to all positive integers only helps us reach the final conclusions as immediate consequences. For any given data set of finite size, the number of mixture components required to approximate a target density will always be bounded above by the number of

latent or observed variables generated by the target density. For most practical applications the required number would actually be much smaller than the number of variables generated by the target. Even if one applies mixture models that a-priori allow potentially infinitely many mixture components, the posterior will essentially concentrate on a finite set comprising moderately small positive integers. This means that for all practical purposes solutions based on finite mixture models with fixed but sufficiently large number of mixture components will essentially be as robust as solutions based on their infinite or varying dimensional counterparts while at the same time being significantly less burdensome from a computational viewpoint. The requirement that the priors on the number of mixture components assign positive mass on *all* positive integers may thus be relegated to the requirement that the priors assign positive mass on sets of the form $\{1, \ldots, K\}$, where K is sufficiently large. Posterior computation for such models might be even much more intensive and complex requiring reversible jump moves. Since a mixture model with K components is at least as flexible as a model with $(K-1)$ components, properties of overfitted mixture models discussed in Section C.1.2 allow us to adopt a much simpler strategy. We can simply keep the number of mixture components fixed at sufficiently large values for all MCMC iterations. Carefully chosen priors for the mixture probabilities then result in a posterior that concentrates in regions favoring empty redundant components, essentially eliminating the need to assign any priors on the number of mixture components. We will still need some mechanism, preferably an automated and data adaptive one, to determine what values of K would be sufficiently large. This issue is discussed in the section on hyper-parameter choices in the Appendix.

The discussions of Section C.1 suggest that finite mixture models with sufficiently large number of mixture components and carefully chosen priors for the mixture probabilities can essentially retain the major advantages of infinite dimensional alternatives including flexibility, automated model averaging and model selection while at the same time being computationally much less burdensome, making them our preferred choice for complicated high dimensional problems.

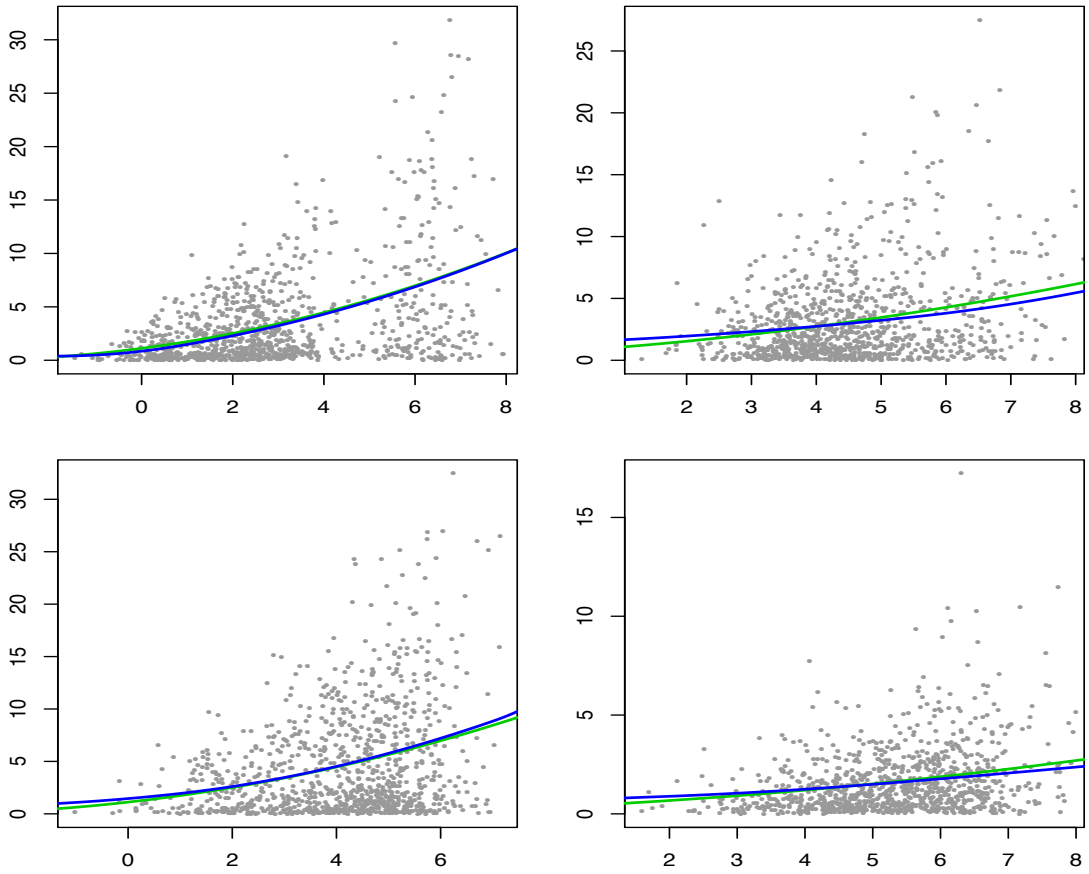

Figure C.1: Results for the variance functions $s^2(X)$ produced by the univariate density deconvolution method for each component of the vector of interest $\mathbf{X}$ for the conditionally heteroscedastic error distribution $f^2_{\boldsymbol{\epsilon}}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and component specific covariance matrices with autoregressive structure (AR). The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets for the MIW (mixtures with inverse Wishart priors) method. For each component of $\mathbf{X}$, the true variance function is $s^2(X) = (1 + X/4)^2$. See Section 4.4.2 and Section 4.7 for additional details. In each panel, the true (lighter shaded lines) and the estimated (darker shaded lines) variance functions are superimposed over a plot of subject specific sample means vs subject specific sample variances. The figure is in color in the electronic version of this dissertation.

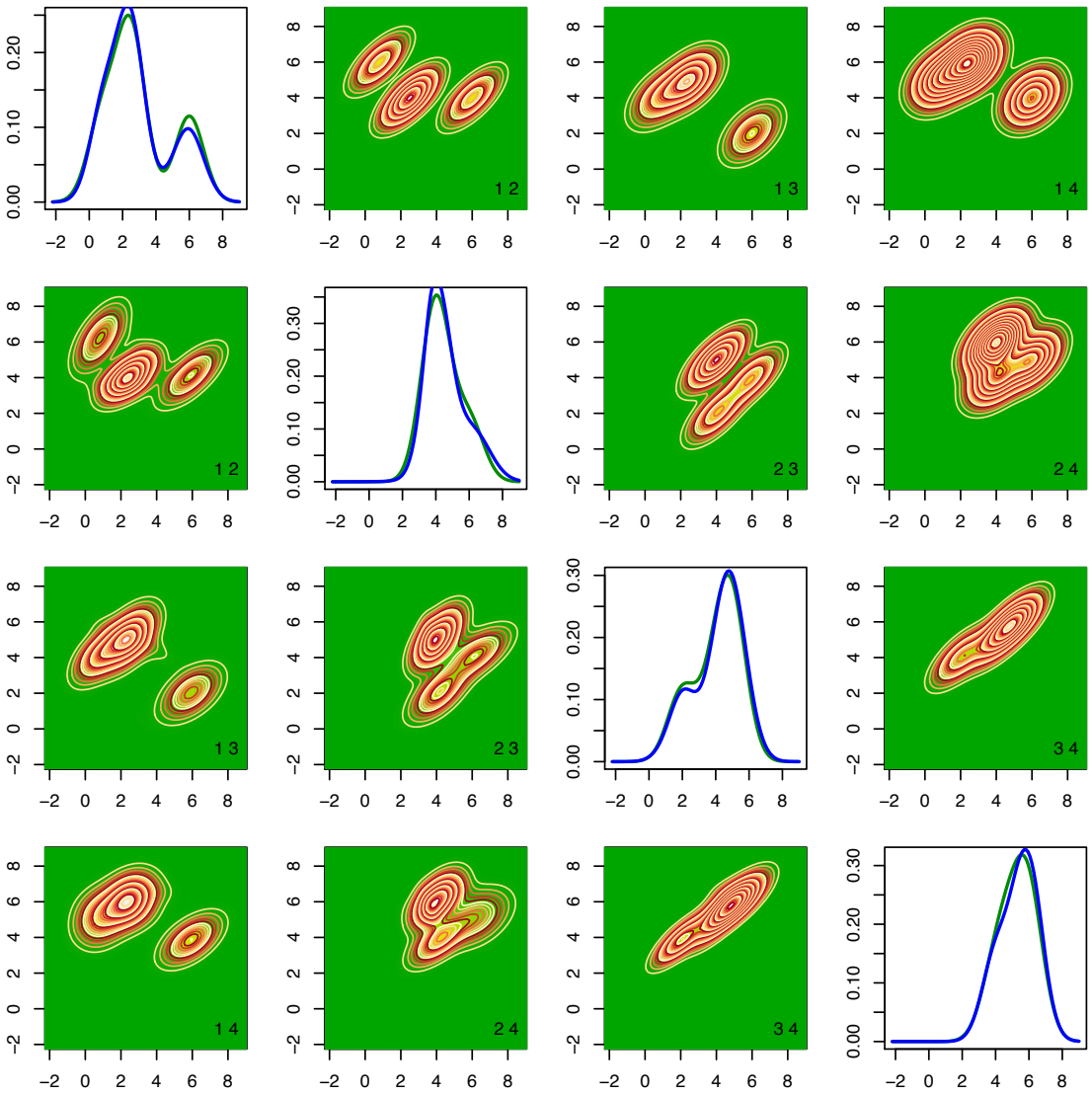

Figure C.2: Results for the density of interest $f_{\mathbf{X}}$ produced by the MIW (mixtures with inverse Wishart priors) method for the conditionally heteroscedastic error distribution $f^2_{\boldsymbol{\epsilon}}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and component specific covariance matrices with autoregressive structure (AR). The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets. See Section 4.8 for additional details. The upper triangular panels show the contour plots of the true two dimensional marginal densities. The lower triangular diagonally opposite panels show the corresponding estimates. The diagonal panels show the true (lighter shaded lines) and the estimated (darker shaded lines) one dimensional marginals. The figure is in color in the electronic version of this dissertation.

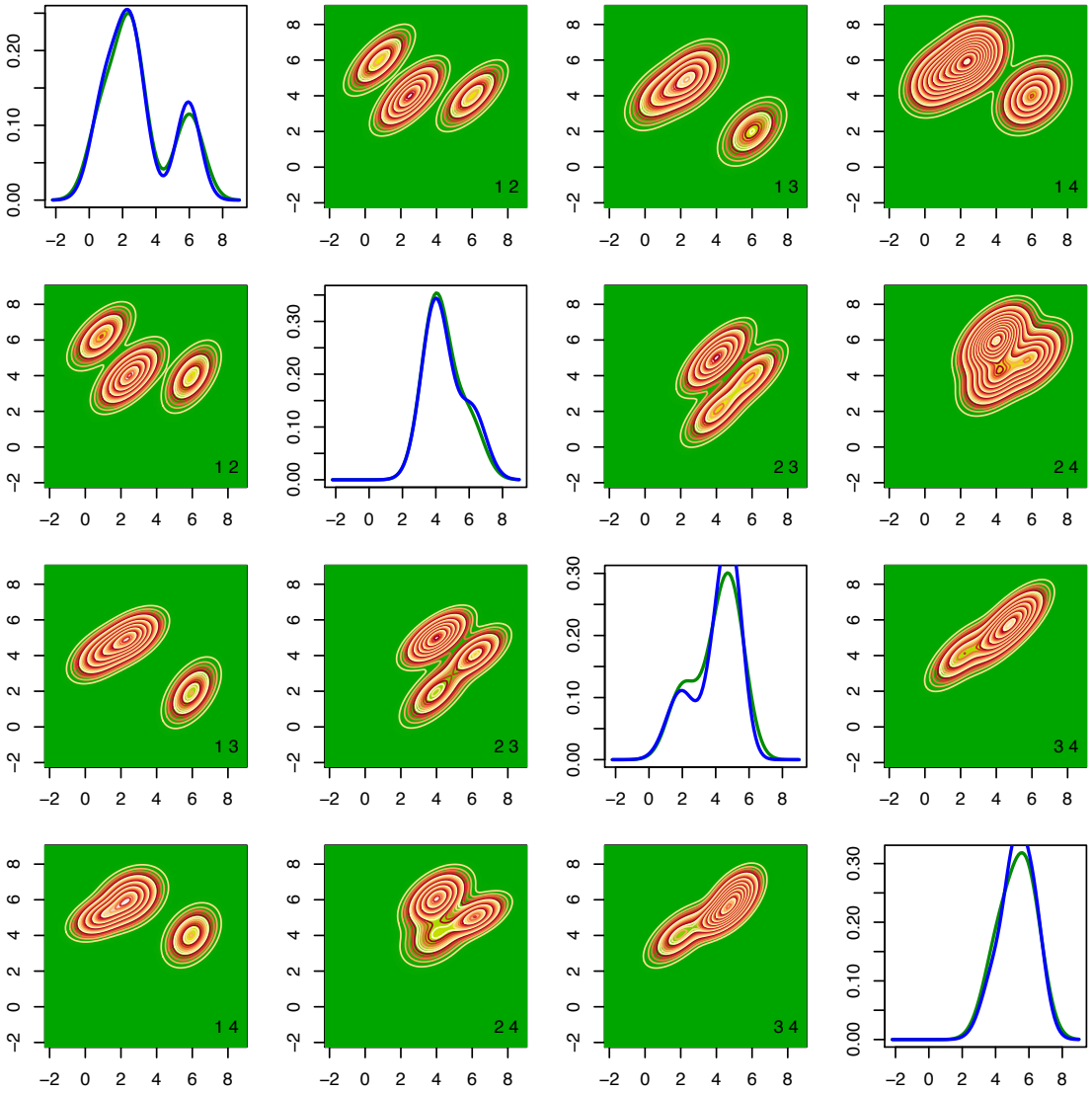

Figure C.3: Results for the density of interest $f_{\mathbf{X}}$ produced by the MLFA (mixtures of latent factor analyzers) method for the conditionally heteroscedastic error distribution $f^2_{\boldsymbol{\epsilon}}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and component specific covariance matrices with autoregressive structure (AR). The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets. See Section 4.8 for additional details. The upper triangular panels show the contour plots of the true two dimensional marginal densities. The lower triangular diagonally opposite panels show the corresponding estimates. The diagonal panels show the true (lighter shaded lines) and the estimated (darker shaded lines) one dimensional marginals. The figure is in color in the electronic version of this dissertation.

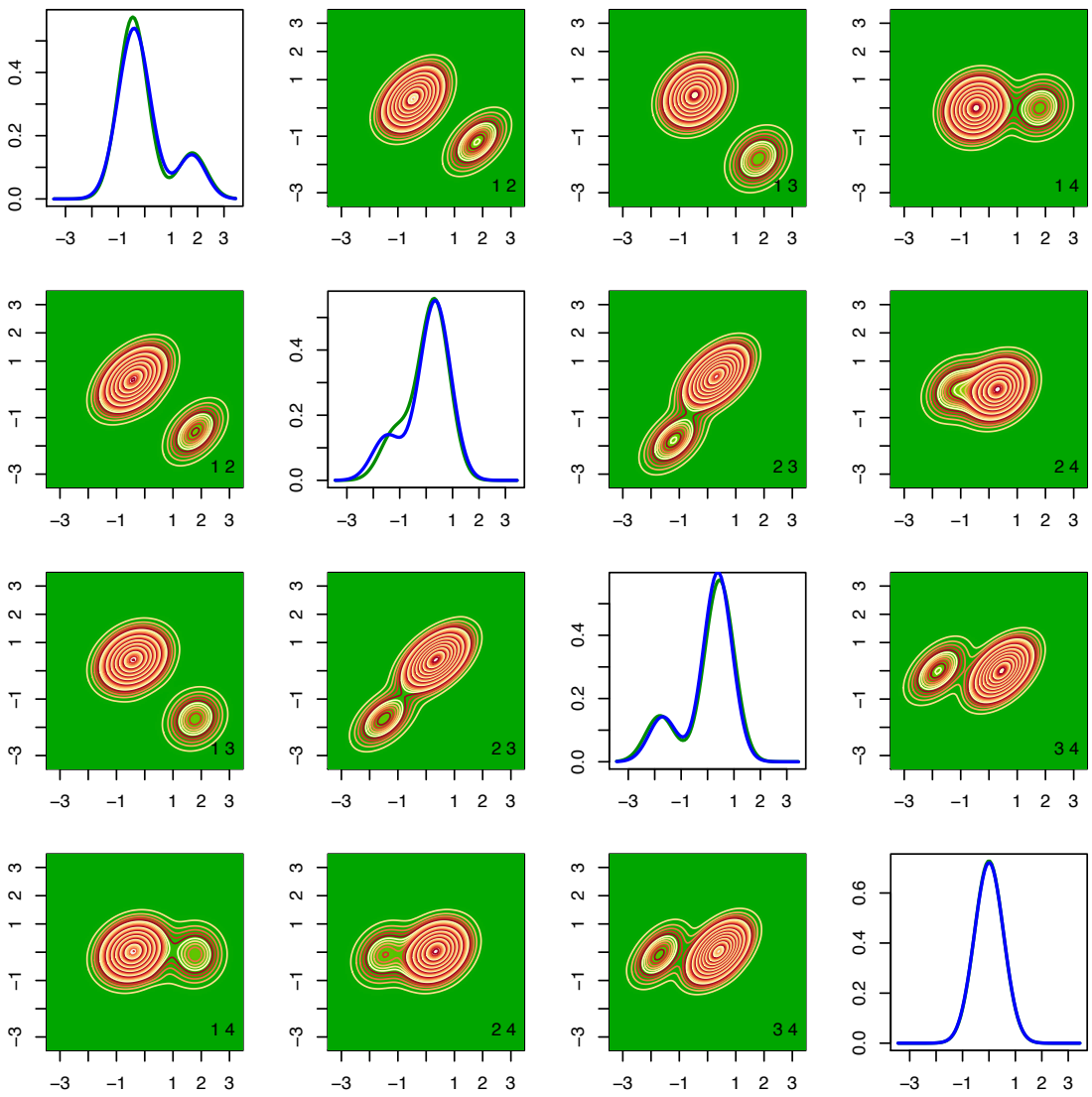

Figure C.4: Results for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$ produced by the MIW (mixtures with inverse Wishart priors) method for the conditionally heteroscedastic error distribution $f^2_{\boldsymbol{\epsilon}}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and component specific covariance matrices with autoregressive structure (AR). The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets. See Section 4.8 for additional details. The upper triangular panels show the contour plots of the true two dimensional marginal densities. The lower triangular diagonally opposite panels show the corresponding estimates. The diagonal panels show the true (lighter shaded lines) and the estimated (darker shaded lines) one dimensional marginals. The figure is in color in the electronic version of this dissertation.

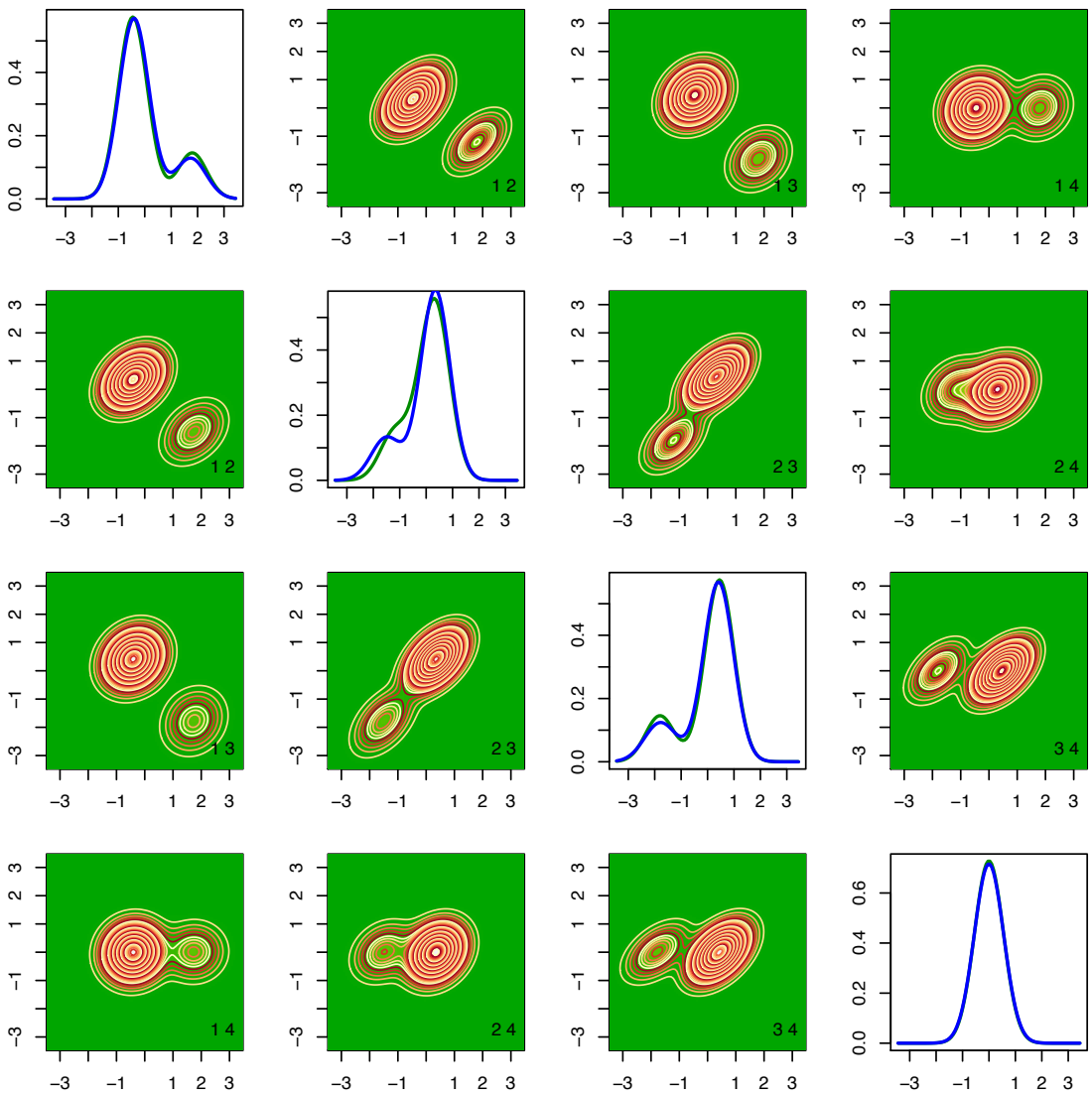

Figure C.5: Results for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$ produced by the MLFA (mixtures of latent factor analyzers) method for the conditionally heteroscedastic error distribution $f^2_{\boldsymbol{\epsilon}}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and component specific covariance matrices with autoregressive structure (AR). The results correspond to the data set that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets. See Section 4.8 for additional details. The upper triangular panels show the contour plots of the true two dimensional marginal densities. The lower triangular diagonally opposite panels show the corresponding estimates. The diagonal panels show the true (lighter shaded lines) and the estimated (darker shaded lines) one dimensional marginals. The figure is in color in the electronic version of this dissertation.

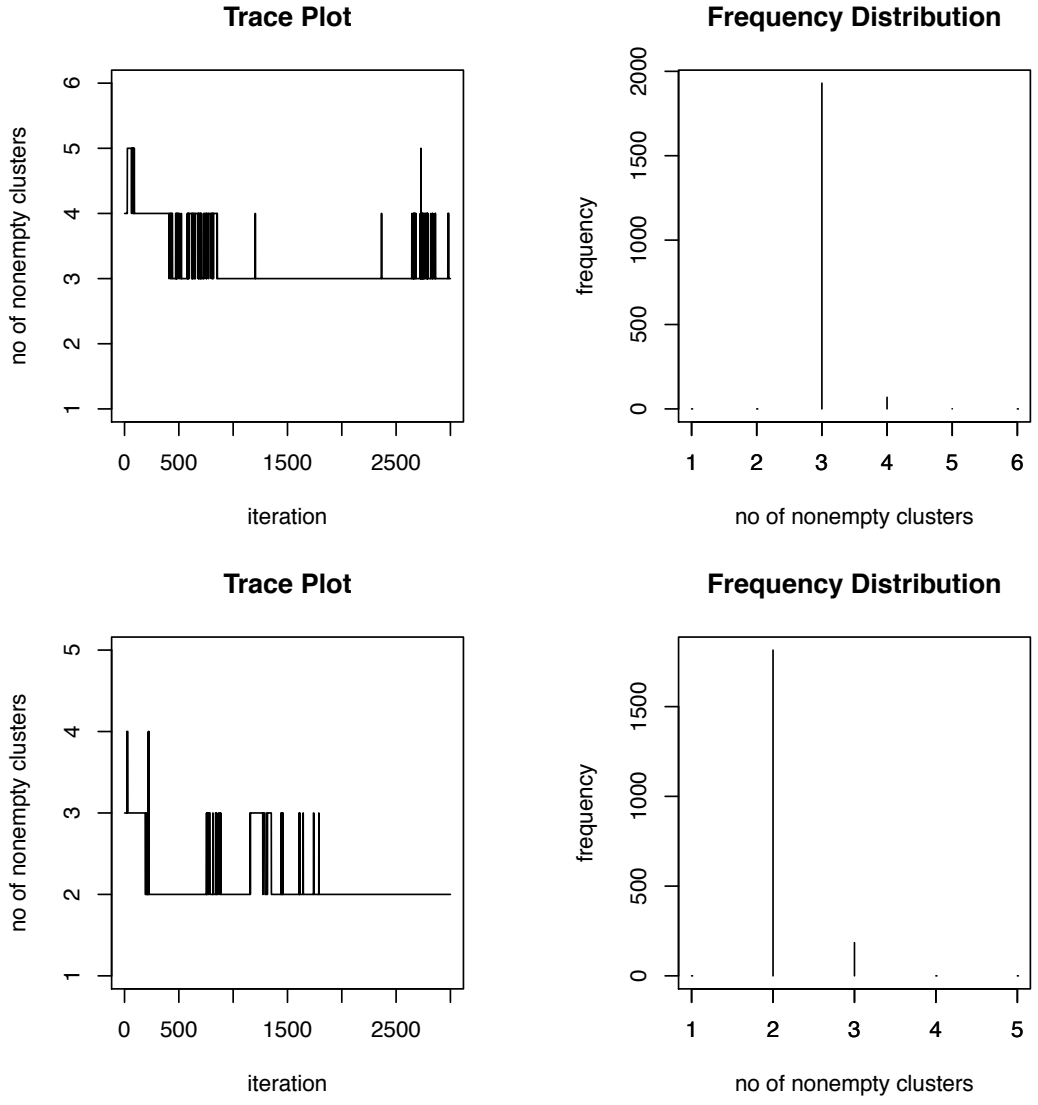

Figure C.6: Trace plots and frequency distributions of the number of nonempty clusters produced by the MIW (mixtures with inverse Wishart priors) method for the conditionally heteroscedastic error distribution $f^2_{\boldsymbol{\epsilon}}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and component specific covariance matrices with autoregressive structure (AR). See Section 4.8 for additional details. The results correspond to the simulation instance that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets, when the number of mixture components for both the density of interest and the density of scaled errors were kept fixed at $K_{\mathbf{X}} = 6$ and $K_{\boldsymbol{\epsilon}} = 5$. The upper panels are for the density of interest $f_{\mathbf{X}}$ and the lower panels are for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$. The true number of mixture components were $K_{\mathbf{X}} = 3$ and $K_{\boldsymbol{\epsilon}} = 3$. As can be seen from Figure C.4, a mixture model with 2 nonempty clusters can approximate the true density of the scaled errors well.

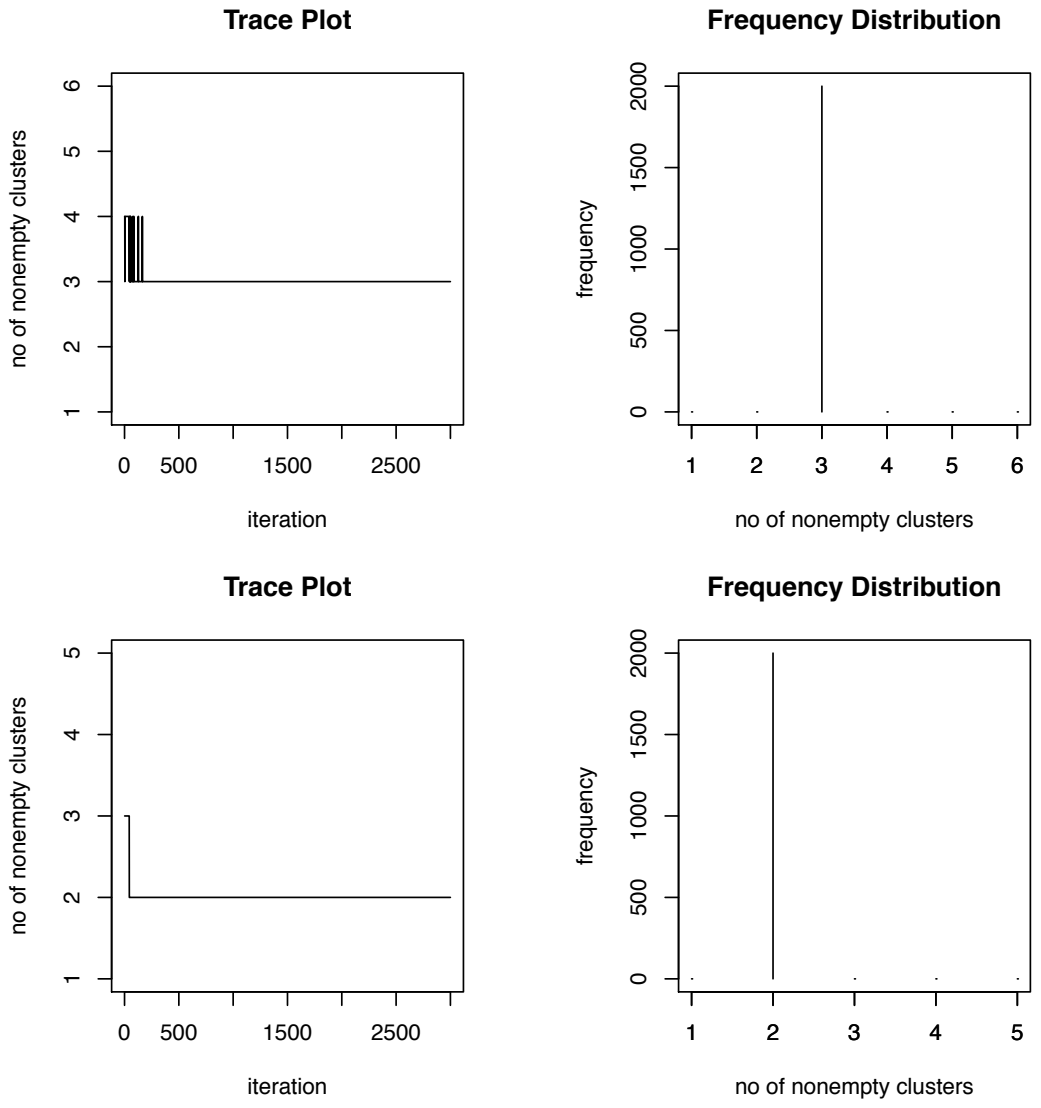

Figure C.7: Trace plots and frequency distributions of the number of nonempty clusters produced by the MLFA (mixtures of latent factor analyzers) method for the conditionally heteroscedastic error distribution $f^2_{\boldsymbol{\epsilon}}$ with sample size $n = 1000$, $m_i = 3$ replicates for each subject and component specific covariance matrices with autoregressive structure (AR). See Section 4.8 for additional details. The results correspond to the simulation instance that produced the median of the estimated integrated squared errors (ISE) out of a total of 100 simulated data sets, when the number of mixture components for both the density of interest and the density of scaled errors were kept fixed at $K_{\mathbf{X}} = 6$ and $K_{\boldsymbol{\epsilon}} = 5$. The upper panels are for the density of interest $f_{\mathbf{X}}$ and the lower panels are for the density of the scaled errors $f_{\boldsymbol{\epsilon}}$. The true number of mixture components were $K_{\mathbf{X}} = 3$ and $K_{\boldsymbol{\epsilon}} = 3$. As can be seen from Figure C.5, a mixture model with 2 nonempty clusters can approximate the true density of the scaled errors well.

APPENDIX D

MODEL FLEXIBILITY

In this appendix we theoretically investigate the flexibility of the regression and the deconvolution models proposed in this thesis. In Section D.1 of this Appendix, we first study flexibility of the regression model proposed in Section 3. Similar results for the deconvolution model of Section 2 are obtained as trivial special cases. Analogous results for multivariate deconvolution models of Section 4 whose proofs require nontrivial modifications due to the multivariate set up and the use of sophisticated priors are presented separately in Section D.3 of this Appendix.

D.1 Flexibility of the Univariate Deconvolution and Regression Models of Section 2 and Section 3

Let the generic notation Π denote a prior on some class of random functions. Also let $\mathcal{T}$ denote the target class of functions to be modeled by Π. The support of Π throws light on the flexibility of Π. For Π to be a flexible prior, one would expect that $\mathcal{T}$ or a large subset of $\mathcal{T}$ would be contained in the support of Π.

For investigating the flexibility of priors for density functions, a relevant concept is that of Kullback-Leibler (KL) support. The KL divergence between two densities f_0 and f, denoted by $d_{KL}(f_0, f)$, is defined as $d_{KL}(f_0, f) = \int f_0(Z) \log \{f_0(Z)/f(Z)\} dZ$. Let Π_f denote a prior assigned to a random density f. A density f_0 is said to belong to the KL support of Π_f if $\Pi_f\{f : d_{KL}(f_0, f) < \delta\} > 0 \; \forall \delta > 0$. The class of densities in the KL support of Π_f is denoted by $KL(\Pi_f)$.

We assume f_X to have support on a closed interval $[A, B]$. Let $\mathcal{F}_X$ denote the set of all densities on $[A, B]$, the target class of densities to be modeled. Also let $\widetilde{\mathcal{F}}_X \subseteq \mathcal{F}_X$ denote the class of densities f_{0X} that satisfy the following regularity conditions.

Conditions 2. *1. f_{0X} is continuous, nowhere zero on $[A, B]$ and is bounded above by some $M < \infty$. 2. $|\int_A^B f_{0X}(Z) \; log \; \{f_{0X}(Z)/\inf_{t\in[A,B]\cap[Z-\delta,Z+\delta]} f_{0X}(t)\} dZ| < \infty$ for some $\delta > 0$.*

Let Π_X denote the prior for the unknown density of X defined in Section 3.2.1. The following lemma is obtained along the lines of Thereom 3.2 of Tokdar (2006) with minor adjustments in the proof for compact support of X.

Lemma 2. $\widetilde{\mathcal{F}}_X \subseteq KL(\Pi_X)$.

Next we consider the model for scaled errors. Let $\mathcal{F}_\epsilon$ denote the set of densities on $\mathbb{R}$ that have mean zero. Also let $\widetilde{\mathcal{F}}_\epsilon \subseteq \mathcal{F}_\epsilon$ denote the set of densities $f_{0\epsilon}$ that satisfy the following regularity assumptions.

Conditions 3. *1.* $f_{0\epsilon}$ *is continuous, nowhere zero on* $\mathbb{R}$ *and is bounded above by some* $M < \infty$. *2.* $\int Z^2 \; f_{0\epsilon}(Z) \; dZ < \infty$. *3.* $|\int \; f_{0\epsilon}(Z) \; log\{f_{0\epsilon}(Z)\} \; dZ| < \infty$. *4.* $\left|\int \; f_{0\epsilon}(Z) \; log\{f_{0\epsilon}(Z)/\inf_{t\in[Z-\delta,Z+\delta]} f_{0\epsilon}(t)\} \; dZ\right| < \infty$ *for some* $\delta > 0$.

Let Π_ϵ denote the prior for the unknown density of ϵ defined in Section 3.2.2. The following lemma follows from straightforward modifications of the results in Pelenis (2014).

Lemma 3. $\widetilde{\mathcal{F}}_\epsilon \subseteq KL(\Pi_\epsilon)$.

To study flexibility of the prior distribution, it is natural to place tail conditions on the class of densities. Conditions 2 and 3 are natural extensions of similar conditions introduced by Tokdar (2006) and encompass large subclasses of $\mathcal{F}_X$ and $\mathcal{F}_\epsilon$, respectively. Lemma 2 and Lemma 3, therefore, establish the flexibility of the models for f_X and f_ϵ, respectively.

For investigating the flexibility of models for general classes of functions, a relevant concept is that of sup norm support. The sup norm distance between two functions g_0 and g, denoted by $||g_0 - g||_\infty$, is defined as $||g_0 - g||_\infty = \sup_Z |g_0(Z) - g(Z)|$. Let Π_g denote a prior assigned to a random function g. A function g_0 is said to belong to the sup norm support of Π_g if $\Pi_g(g : ||g_0 - g||_\infty < \delta) > 0 \; \forall \delta > 0$. The class of functions in the sup norm support of Π_g is denoted by $SN(\Pi_g)$.

Let $\mathcal{C}[A, B]$ denote the set of continuous functions from $[A, B]$ to $\mathbb{R}$. Also let $\mathcal{C}_{+1}[A, B] = \{v : v \in \mathcal{C}[A, B], v > 0, v(t_{j_0}) = 1\}$.

Let Π_R denote the prior on the regression function r defined in Section 3.2.3 and Π_V denote the prior on the variance function v defined in Section 3.2.2. The first part of the following lemma follows directly from the results on page 147 of de Boor (2000). The proof of the second part is given in Appendix D.2.

Lemma 4. *1.* $\mathcal{C}[A, B] \subseteq SN(\Pi_R)$. *2.* $\mathcal{C}_{+1}[A, B] \subseteq SN(\Pi_V)$.

For a given X, let $\Pi_{Y|X}$ denote the induced prior for $f_{Y|X}$. When X is not specified, let $\Pi_{Y|\bullet}$ denote the prior for the conditional density of Y induced by Π_R, Π_ϵ and Π_V. Define $\Pi_{\mathbf{W}_{1:m}|X}$ and $\Pi_{\mathbf{W}_{1:m}|\bullet}$ in similar fashion.

Define $\widetilde{\mathcal{F}}_{Y|X} = \{f_{0Y|X} : f_{0Y|X}(Y) = v_{0Y}^{-1/2}(X) f_{0\epsilon_Y}[v_{0Y}^{-1/2}(X)\{Y - r_0(X)\}], r_0 \in \mathcal{C}[A,B], v_{0Y} \in \mathcal{C}_{+1}[A,B], f_{0e_Y} \in \widetilde{\mathcal{F}}_\epsilon\}$ and $\widetilde{\mathcal{F}}_{Y|\bullet} = \{f_{0Y|\bullet} : f_{0Y|\bullet} = f_{0Y|X} \in \widetilde{\mathcal{F}}_{Y|X}\ \forall X \in [A,B]\}$. Let $\widetilde{\mathcal{F}}_{\mathbf{W}_{1:m}|X} = \{f_{0\mathbf{W}_{1:m}|X} : v_{0W} \in \mathcal{C}_{+1}[A,B], f_{0e_W} \in \widetilde{\mathcal{F}}_\epsilon, f_{0\mathbf{W}_{1:m}|X}(\mathbf{W}_{1:m}) = v_{0W}^{-m/2}(X) \prod_{j=1}^m f_{0\epsilon_W}\{v_{0W}^{-1/2}(X)(W_j - X)\}\}$ and $\widetilde{\mathcal{F}}_{\mathbf{W}_{1:m}|\bullet} = \{f_{0\mathbf{W}_{1:m}|\bullet} : f_{0\mathbf{W}_{1:m}|\bullet} = f_{0\mathbf{W}_{1:m}|X} \in \widetilde{\mathcal{F}}_{\mathbf{W}_{1:m}|X}\ \forall X \in [A,B]\}$. The following lemma tells us about the support of the conditional densities $f_{Y|X}$ and $f_{\mathbf{W}_{1:m}|X}$ implied by our models.

Lemma 5. *1.* $\widetilde{\mathcal{F}}_{Y|X} \subseteq KL(\Pi_{Y|X})$ *for any* $X \in [A,B]$.
2. For any $f_{0Y|\bullet} \in \widetilde{\mathcal{F}}_{Y|\bullet}$, $\Pi_{Y|\bullet}\{\sup_{X\in[A,B]} d_{KL}(f_{0Y|X}, f_{Y|X}) < \delta\} > 0\ \forall \delta > 0$.
3. $\widetilde{\mathcal{F}}_{\mathbf{W}_{1:m}|X} \subseteq KL(\Pi_{\mathbf{W}_{1:m}|X})$ *for any* $X \in [A,B]$.
4. For any $f_{0\mathbf{W}_{1:m}|\bullet} \in \widetilde{\mathcal{F}}_{\mathbf{W}_{1:m}|\bullet}$, $\Pi_{\mathbf{W}_{1:m}|V}\{\sup_{X\in[A,B]} d_{KL}(f_{0\mathbf{W}_{1:m}|X}, f_{\mathbf{W}_{1:m}|X}) < \delta\} > 0\ \forall \delta > 0$.

Finally, we investigate the support of the induced prior for $f_{Y,\mathbf{W}_{1:m}}$ which we denote by $\Pi_{Y,\mathbf{W}_{1:m}}$. The support of $\Pi_{Y,\mathbf{W}_{1:m}}$ tells us about the types of likelihood functions the models can approximate.

Let $f_{conv}(Y, \mathbf{W}_{1:m} \mid f_X, r, v_Y, v_W, f_{\epsilon_Y}, f_{\epsilon_W})$ denote the density obtained by the convolution $\int v_Y^{-1/2}(X) v_W^{-m/2}(X) f_{\epsilon_Y}[v_Y^{-1/2}(X)\{Y - r(X)\}] \prod_{j=1}^m f_{\epsilon_W}\{v_W^{-1/2}(X)(W_j - X)\} f_X(X) dX$. Define $\widetilde{\mathcal{F}}_{Y,\mathbf{W}_{1:m}} = \{f_{0Y,\mathbf{W}_{1:m}} : f_{0Y,\mathbf{W}_{1:m}}(Y, \mathbf{W}_{1:m}) = f_{conv}(Y, \mathbf{W}_{1:m} \mid f_{0X}, r_0, v_{0Y}, v_{0W}, f_{0\epsilon_Y}, f_{0\epsilon_W});\ f_{0X} \in \widetilde{\mathcal{F}}_X; r_0 \in \mathcal{C}[A,B]; v_{0Y}, v_{0W} \in \mathcal{C}_{+1}[A,B]; f_{0e_Y}, f_{0e_W} \in \widetilde{\mathcal{F}}_\epsilon\}$.

Theorem 1. $\widetilde{\mathcal{F}}_{Y,\mathbf{W}_{1:m}} \subseteq KL(\Pi_{Y,\mathbf{W}_{1:m}})$.

The proofs of Lemma 5 and Theorem 1 are given in Appendix D.2. The proof of Lemma 4 require that the priors Π_R and Π_V allow the number of knot points to vary. Posterior inference for such models via Markov chain Monte Carlo techniques will be highly computation intensive. Our experience with simulation studies suggests that, provided the true regression and variance functions are smooth and $15 \leq K_R \leq 25$ and $5 \leq K \leq 10$, the results are insensitive to the number of knot points used. To reduce computational costs, in the simulation studies and the real world application presented in this dissertation, the number of knots are, therefore, kept fixed.

D.2 Proofs of the Theoretical Results of Appendix D.1

D.2.1 Proof of Lemma 4

Given q, let Π_q denote a prior on $\mathbb{N}_q = \{q+1, q+2, \dots\}$ such that $\Pi_q(J) > 0 \ \forall J \in \mathbb{N}_q$. Let $||\cdot||_2$ denote the Euclidean norm. Given $J \sim \Pi_q$, also let $\Pi_{\beta|J}$ be a prior on $\mathbb{R}^{+J}$ such that $\Pi_{\beta|J}\{N_\delta(\boldsymbol{\beta}_0)\} > 0$ for any $\delta > 0$ and any $\boldsymbol{\beta}_0 \in \mathbb{R}^J$, where $N_\delta(\boldsymbol{\beta}_0) = \{\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathbb{R}^J, ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_2 < \delta\}$. Define $\mathcal{S}_{q,J} = \{r_s : r_s = \mathbf{B}_{q,J}\boldsymbol{\beta} = \sum_{j=1}^J b_{q,j}\beta_j \text{ for some } \boldsymbol{\beta} \in \mathbb{R}^J\}$. Then $\Pi_q \times \Pi_{\beta|J}$ is the induced prior on $\mathcal{S}_q = \cup_{J=q+1}^\infty \mathcal{S}_{q,J}$. Define $\psi(v_0, h) = \sup_{X,X' \in [A,B], |X-X'| \le h} |r_0(X) - r_0(X')|$. Let $\lfloor \alpha \rfloor = \min\{n : n \in \mathbb{N}, n \ge \alpha\}$. Using the local support properties of B-splines mentioned, the result page 147 of de Boor (2000) says that, for any $r_0 \in \mathcal{C}[A,B]$,

$$\inf_{r_s \in \mathcal{S}_{q,J}} ||r_0 - r_s||_\infty \le \lfloor (q+1)/2 \rfloor \ \psi(r_0, \Delta_{\max}) \to 0 \ \text{ as } \Delta_{\max} \to 0.$$

Given any $r_0 \in C[A,B]$ and $\delta > 0$, find $J \in \mathbb{N}_q$ and $\boldsymbol{\beta}_0 \in \mathbb{R}^J$ such that $||r_0 - \mathbf{B}_{q,J}\boldsymbol{\beta}_0||_\infty = \inf_{r_s \in \mathcal{S}_{q,J}} ||r_0 - r_s||_\infty < \delta/2$. Next consider a neighborhood $N_\eta(\boldsymbol{\beta}_0)$ such that for any $\boldsymbol{\beta} \in N_\eta(\boldsymbol{\beta}_0)$, we have $||\mathbf{B}_{q,J}\boldsymbol{\beta} - \mathbf{B}_{q,J}\boldsymbol{\beta}_0||_\infty < \delta/2$. Then for any $\boldsymbol{\beta} \in N_\eta(\boldsymbol{\beta}_0)$, we have $||\mathbf{B}_{q,J}\boldsymbol{\beta} - v_0||_\infty \le ||\mathbf{B}_{q,J}\boldsymbol{\beta} - \mathbf{B}_{q,J}\boldsymbol{\beta}_0||_\infty + ||\mathbf{B}_{q,J}\boldsymbol{\beta}_0 - r_0||_\infty < \delta$. Also $\Pi_V(||r - r_0||_\infty < \delta) \ge \Pi_q(J)\ \Pi_{\beta|J}\{N_\eta(\boldsymbol{\beta}_0)\} > 0$. Part 1 of Lemma 4 then follows as a special case taking $\boldsymbol{\beta} = \boldsymbol{\xi}_R$, Π_q to be the prior on J induced by $p_0(K_R)$, $\Pi_{\beta|J}$ to be the prior on $\boldsymbol{\beta}$ induced by $p_0(\boldsymbol{\xi}_R \mid K_R, \sigma^2_{R,\xi})$, and $\Pi_R = \Pi_q \times \Pi_{\beta|J}$.

To prove part 2 of Lemma 4, let $\mathbb{R}_+ = (0, \infty)$, $\mathcal{C}_+[A,B] = \{g_s : g_s \in \mathcal{C}[A,B], g > 0\}$ and $\mathcal{S}^+_{q,J} = \{g : g = \mathbf{B}_{q,J}\boldsymbol{\beta}_+ \text{ for some } \boldsymbol{\beta}_+ \in \mathbb{R}^J_+\}$. Then the result of de Boor (2000) can be easily modified to prove that, for any $g_0 \in \mathcal{C}_+[A,B]$,

$$\inf_{g_s \in \mathcal{S}^+_{q,J}} ||g_0 - g_s||_\infty \le \lfloor (q+1)/2 \rfloor \ \psi(g_0, \Delta_{\max}) \to 0 \ \text{ as } \Delta_{\max} \to 0.$$

Now, let $\mathbb{R}^J_{+1} = \{\boldsymbol{\beta}_{+1} : \boldsymbol{\beta}_{+1} \in \mathbb{R}^J_+, \sum_{j=(j_0-q)}^{(j_0-1)} b_{q,j}(t_{j_0})\beta_{+1,j} = 1\}$. Given $J \sim \Pi_q$, let $\Pi^{+1}_{\beta|J}$ be a prior on $\mathbb{R}^J_{+1}$ such that $\Pi^{+1}_{\beta|J}\{N_\delta(\boldsymbol{\beta}_{+1,0})\} > 0$ for any $\delta > 0$ and any $\boldsymbol{\beta}_{+1,0} \in \mathbb{R}^J_{+1}$. Define $\mathcal{S}^{+1}_{q,J} = \{v_s : v_s = \mathbf{B}_{q,J}\boldsymbol{\beta}_{+1}, \text{ for some } \boldsymbol{\beta}_{+1} \in \mathbb{R}^J_{+1}\}$. Then $\Pi_q \times \Pi^{+1}_{\beta|J}$ is the induced prior on $\mathcal{S}^{+1}_q = \cup_{J=q+1}^\infty \mathcal{S}^{+1}_{q,J}$. Given $v_s = \mathbf{B}_{q,J}\boldsymbol{\beta}_{+1} \in \mathcal{S}^{+1}_{q,J}$, define $\mathcal{S}^+_{q,J}(v_s)$ comprising functions $g_s = \mathbf{B}_{q,J}\boldsymbol{\beta}_+ \in \mathcal{S}^+_{q,J}$, where $\beta_{+,j} = \beta_{+1,j}$ for $j \ne (j_0 - q), \dots, (j_0 - 1)$, but $\beta_{+,j}$ can be different from $\beta_{+1,j}$ for $j = (j_0 - q), \dots, (j_0 - 1)$. Note that

$\cup_{v_s \in \mathcal{S}^{+1}_{q,J}} \mathcal{S}^{+}_{q,J}(v_s) = \mathcal{S}^{+}_{q,J}$. Also note that $v_s(X) = g_s(X)\ \forall X \in [A, t_{j_0-q}] \cup [t_{j_0+q}, B]$ and $\forall g_s \in \mathcal{S}^{+}_{q,J}(v_s)$. Therefore, for any $v_0 \in \mathcal{C}_{+1}[A, B]$,

$$\begin{aligned} \sup_{X \in [A, t_{j_0-q}] \cup [t_{j_0+q}, B]} |v_0(X) - v_s(X)| &= \sup_{X \in [A, t_{j_0-q}] \cup [t_{j_0+q}, B],\ g_s \in \mathcal{S}^{+}_{q,J}(v_s)} |v_0(X) - g_s(X)| \\ &\leq \inf_{g_s \in \mathcal{S}^{+}_{q,J}(v_s)} ||v_0 - g_s||_\infty. \end{aligned}$$

Also, by definition, $|v_0(X) - v_s(X)| \to 0$ as $X \to t_{j_0}$, and for fixed q, as $K \to \infty$, $\Delta_{\max} \to 0$, $t_{j_0} \geq t_{j_0-q} \geq (t_{j_0} - q\Delta_{\max}) \to t_{j_0}$ and $t_{j_0} \leq t_{j_0+q} \leq (t_{j_0} + q\Delta_{\max}) \to t_{j_0}$. Therefore, given any $\delta > 0$, since $(v_0 - v_s)$ is uniformly continuous on $[A, B]$, for K sufficiently large,

$$\sup_{X \in [t_{j_0-q}, t_{j_0+q}]} |v_0(X) - v_s(X)| < \delta.$$

Combining, we have, for any given $\delta > 0$, for K sufficiently large,

$$\begin{aligned} ||v_0 - v_s||_\infty &\leq \max\{\delta, \inf_{g_s \in \mathcal{S}^{+}_{q,J}(v_s)} ||v_0 - g_s||_\infty\}. \\ \Rightarrow \inf_{v_s \in \mathcal{S}^{+1}_{q,J}} ||v_0 - v_s||_\infty &\leq \max\{\delta, \inf_{v_s \in \mathcal{S}^{+1}_{q,J}} \inf_{g_s \in \mathcal{S}^{+}_{q,J}(v_s)} ||v_0 - g_s||_\infty\} \\ &= \max\{\delta, \inf_{g_s \in \mathcal{S}^{+}_{q,J}} ||v_0 - g_s||_\infty\} = \delta. \end{aligned}$$

Taking $\boldsymbol{\beta}^+ = \exp(\boldsymbol{\xi})$, an argument along the lines of the proof of part 1 then completes the proof of part 2.

D.2.2 Proof of Lemma 5

Recall the specification of f_ϵ given by (3.6). Letting $(p, \widetilde{\mu}, \sigma_1^2, \sigma_2^2)^{\mathrm{T}} = \boldsymbol{\theta}$ and $P_\epsilon(\boldsymbol{\theta}) = \sum_{k=1}^{\infty} \pi_{\epsilon k} \delta_{\boldsymbol{\theta}_k}(\boldsymbol{\theta})$, where δ_θ denotes a point mass at θ, we have $f_\epsilon(\epsilon) = \int f_{c\epsilon}(\epsilon \mid \boldsymbol{\theta}) dP_\epsilon(\boldsymbol{\theta})$. The following two lemmas prove some properties of P_ϵ and f_ϵ.

Lemma 6. *1.* $\int \max\{\sigma_1^{-1}, \sigma_2^{-1}\}\ dP_\epsilon(\boldsymbol{\theta}) < \infty$ *a.s.*
2. $\int |log\ p|\ dP_\epsilon(\boldsymbol{\theta}) < \infty$ *a.s.*
3. $\int |log\ \sigma_1|\ dP_\epsilon(\boldsymbol{\theta}) < \infty$ *a.s.*
4. $\int \sigma_1^{-2}\ dP_\epsilon(\boldsymbol{\theta}) < \infty$ *a.s.*
5. $\int \sigma_1^{-2}\ c_1^2(p)\ \widetilde{\mu}^2\ dP_\epsilon(\boldsymbol{\theta}) < \infty$ *a.s.*

Proof.

1. $\int \max\{\sigma_1^{-1}, \sigma_2^{-1}\}\ dP_\epsilon(\boldsymbol{\theta}) = \int \max\{\tau_1^{1/2}, \tau_2^{1/2}\}\mathrm{Gamma}(\tau_1 \mid a_\epsilon, b_\epsilon)\mathrm{Gamma}(\tau_2 \mid a_\epsilon, b_\epsilon)d\tau_1 d\tau_2$
   $= \int_{\tau_1>\tau_2} \tau_1^{1/2}\mathrm{Gamma}(\tau_1 \mid a_\epsilon, b_\epsilon)\mathrm{Gamma}(\tau_2 \mid a_\epsilon, b_\epsilon)d\tau_1 d\tau_2$
   $+ \int_{\tau_2>\tau_1} \tau_2^{1/2}\mathrm{Gamma}(\tau_1 \mid a_\epsilon, b_\epsilon)\mathrm{Gamma}(\tau_2 \mid a_\epsilon, b_\epsilon)d\tau_1 d\tau_2$
   $\leq 2\int \tau^{1/2}\mathrm{Gamma}(\tau \mid a_\epsilon, b_\epsilon)d\tau$
   $= 2\int_{\tau<1} \tau^{1/2}\mathrm{Gamma}(\tau \mid a_\epsilon, b_\epsilon)d\tau + 2\int_{\tau>1} \tau^{1/2}\mathrm{Gamma}(\tau \mid a_\epsilon, b_\epsilon)d\tau$
   $\leq 2\int_{\tau<1} \mathrm{Gamma}(\tau \mid a_\epsilon, b_\epsilon)d\tau + 2\int_{\tau>1} \tau\mathrm{Gamma}(\tau \mid a_\epsilon, b_\epsilon)d\tau < \infty.$

2. $\int |\log\ p|\ dP_\epsilon(\boldsymbol{\theta}) = -\int \log\ p\ \mathrm{Unif}(p \mid [0,1])\ dp = 1$, since $-\log\ p$ is distributed as a standard exponential random variable.

3. $\int |\log\ \sigma_1|\ dP_\epsilon(\boldsymbol{\theta}) = (1/2)\int_{\sigma_1<1} \log\ \sigma_1^{-2}\ \text{Inv-Gamma}(\sigma_1^2 \mid a_\epsilon, b_\epsilon)$
   $+ (1/2)\int_{\sigma_1>1} \log\ \sigma_1^2\ \text{Inv-Gamma}(\sigma_1^2 \mid a_\epsilon, b_\epsilon)\ d\sigma_1^2$
   $\leq (1/2)\int_{\sigma_1<1} \sigma_1^{-2}\ \text{Inv-Gamma}(\sigma_1^2 \mid a_\epsilon, b_\epsilon)\ d\sigma_1^2 + (1/2)\int_{\sigma_1>1} \sigma_1^2\ \text{Inv-Gamma}(\sigma_1^2 \mid a_\epsilon, b_\epsilon)\ d\sigma_1^2 \leq (1/2)\int \tau\ \mathrm{Gamma}(\tau \mid a_\epsilon, b_\epsilon)\ d\tau + (1/2)\int \sigma_1^2\ \text{Inv-Gamma}(\sigma_1^2 \mid a_\epsilon, b_\epsilon)\ d\sigma_1^2 < \infty$, whenever $a_\epsilon > 1$.

4. $\int \sigma_1^{-2}\ dP_\epsilon(\boldsymbol{\theta}) = \int \tau\ \mathrm{Gamma}(\tau \mid a_\epsilon, b_\epsilon)d\tau < \infty.$

5. Note that $c_1^2(p) = (1-p)^2/\{p^2 + (1-p)^2\} \leq \{p^2 + (1-p)^2\}^{-1} \leq 2$. Therefore, $\int \sigma_1^{-2}\ c_1^2(p)\ \widetilde{\mu}^2\ dP_\epsilon(\boldsymbol{\theta}) \leq \int 2\tau\widetilde{\mu}^2\ \mathrm{Gamma}(\tau \mid a_\epsilon, b_\epsilon)\ \mathrm{Normal}(\widetilde{\mu} \mid 0, \sigma_{\widetilde{\mu}}^2)d\tau d\widetilde{\mu} < \infty.$

Lemma 7. *Let $f_{0\epsilon} \in \widetilde{\mathcal{F}}_\epsilon$ and $f_\epsilon \sim \Pi_\epsilon$. Then*

$$\lim_{(\psi,\tau)\to(0,1)} \int f_{0\epsilon}(\epsilon)\ log\left[\frac{f_\epsilon(\epsilon)}{\tau^{-1} f_\epsilon\{\tau(\epsilon-\psi)\}}\right]\ d\epsilon = 0.$$

Proof. We have $\tau^{-1} f_{c\epsilon}\{\tau(\epsilon-\psi)\} \to f_{c\epsilon}(\epsilon)$ as $(\psi,\tau)\to(0,1)$. Since $\tau \to 1$, without loss of generality, we may assume $\tau > 1/2$. Define $c = \int \max\{\sigma_1^{-1}, \sigma_2^{-1}\} dP_\epsilon(\boldsymbol{\theta}) < \infty$ a.s. Then $\int \tau^{-1} f_{c\epsilon}\{\tau(\epsilon-\psi) \mid \boldsymbol{\theta}\} dP_\epsilon(\boldsymbol{\theta}) \le \int 2(2\pi)^{-1/2}\left(\sigma_1^{-1}+\sigma_2^{-1}\right) dP_\epsilon(\boldsymbol{\theta}) < 2c < \infty$. Applying DCT, $\tau^{-1} f_\epsilon\{\tau(\epsilon-\psi)\} \to f_\epsilon(\epsilon)$ as $(\psi,\tau)\to(0,1)$. Therefore, for any $\epsilon \in \mathbb{R}$,

$$\log\left[\frac{f_\epsilon(\epsilon)}{\tau^{-1} f_\epsilon\{\tau(\epsilon-\psi)\}}\right] \to 0 \quad \text{as } (\psi,\tau)\to(0,1).$$

Let $p_1 = p = (1-p_2)$. Then

$$\begin{aligned}|\log\ f_\epsilon\{\tau(\epsilon-\psi)\}| &\le \log(2\pi)^{1/2} \\ &+ \left|\log \int \sum_{k=1}^{2}\left[\frac{p_k}{\sigma_k}\exp\left\{-\frac{1}{2\sigma_k^2}(\tau\epsilon-\tau\psi-\mu_k)^2\right\}\right] dP_\epsilon(\boldsymbol{\theta})\right|.\end{aligned}$$

Also note that

$$\begin{aligned}\int \sum_{k=1}^{2}\left[\frac{p_k}{\sigma_k}\exp\left\{-\frac{1}{2\sigma_k^2}(\tau\epsilon-\tau\psi-\mu_k)^2\right\}\right] dP_\epsilon(\boldsymbol{\theta}) &\le \int \left(p_1\sigma_1^{-1}+p_2\sigma_2^{-1}\right) dP_\epsilon(\boldsymbol{\theta}) \\ &\le \int (p_1+p_2)\max\{\sigma_1^{-1},\sigma_2^{-2}\} dP_\epsilon(\boldsymbol{\theta}) = c.\end{aligned}$$

Therefore, applying Jensen's inequality on $g(Z) = -\log Z$, we have

$$\begin{aligned}&\left|\log \int \sum_{k=1}^{2}\left[\frac{p_k}{\sigma_k}\exp\left\{-\frac{1}{2\sigma_k^2}(\tau\epsilon-\tau\psi-\mu_k)^2\right\}\right] dP_\epsilon(\boldsymbol{\theta})\right| \\ &\quad\le |\log\ c| - \log \int \sum_{k=1}^{2}\left[\frac{p_k}{c\sigma_k}\exp\left\{-\frac{1}{2\sigma_k^2}(\tau\epsilon-\tau\psi-\mu_k)^2\right\}\right] dP_\epsilon(\boldsymbol{\theta}) \\ &\quad\le |\log\ c| + \log\ c - \int \log\left[\sum_{k=1}^{2}\frac{p_k}{\sigma_k}\exp\left\{-\frac{1}{2\sigma_k^2}(\tau\epsilon-\tau\psi-\mu_k)^2\right\}\right] dP_\epsilon(\boldsymbol{\theta}) \\ &\quad\le |\log\ c| + \log\ c - \int \log(p_1/\sigma_1) dP_\epsilon(\boldsymbol{\theta}) + \int \frac{1}{\sigma_1^2}(\tau^2\epsilon^2+\tau^2\psi^2+\mu_1^2) dP_\epsilon(\boldsymbol{\theta}).\end{aligned}$$

Since $(\psi,\tau)\to(0,1)$, without loss of generality we may also assume $\psi^2 < 1$ and

$\tau < 2$. Therefore,

$$|\log\, f_\epsilon(\tau\epsilon)| \leq 2|\log\, c| - \int \log(p/\sigma_1) dP_\epsilon(\boldsymbol{\theta}) + \int \sigma_1^{-2}\{4\epsilon^2 + 4 + c_1^2(p)\ \widetilde{\mu}^2\} dP_\epsilon(\boldsymbol{\theta}).$$

The regularity assumptions on $f_{0\epsilon}$ and Lemma 6 imply that the RHS above is $f_{0\epsilon}$ integrable. The conclusion of Lemma 7 follows from an application of DCT again.

Now to prove Lemma 5, let $f_{U|(\mu,\sigma)}$ denote the density of $U = (\mu + \sigma\epsilon)$. We have $f_{U|(\mu,\sigma)}(U) = \sigma^{-1} f_\epsilon\{(U-\mu)/\sigma\}$. This implies

$$\begin{aligned}
&\int f_{0U|(\mu_0,\sigma_0)}(U) \log \frac{f_{0U|(\mu_0,\sigma_0)}(U)}{f_{U|(\mu,\sigma)}(U)} dU \\
&= \int f_{0U|(\mu_0,\sigma_0)}(U) \log \frac{f_{0U|(\mu_0,\sigma_0)}(U)}{f_{U|(\mu_0,\sigma_0)}(U)} dU + \int f_{0U|(\mu_0,\sigma_0)}(U) \log \frac{f_{U|(\mu_0,\sigma_0)}(U)}{f_{U|(\mu,\sigma)}(U)} dU \\
&= \int f_{0\epsilon}(\epsilon) \log \frac{f_{0\epsilon}(\epsilon)}{f_\epsilon(\epsilon)} d\epsilon + \int f_{0\epsilon}(\epsilon) \log \frac{f_\epsilon(\epsilon)}{(\sigma_0/\sigma)^{-1} f_\epsilon\{\epsilon\sigma_0/\sigma + (\mu_0 - \mu)/\sigma\}} d\epsilon.
\end{aligned}$$

Let $\delta > 0$ be given. By Lemma 3, $\Pi_\epsilon\{f_\epsilon : d_{KL}(f_{0\epsilon}, f_\epsilon) < \delta/2\} > 0$. By Lemma 7, there exists $\eta_\mu > 0$ and $\eta_\sigma > 0$ such that $|\mu_0 - \mu| < \eta_\mu$ and $|\sigma_0 - \sigma| < \eta_\sigma$ implies $\int f_{0\epsilon}(\epsilon)\ \log[f_\epsilon(\epsilon)/[(\sigma_0/\sigma)^{-1} f_\epsilon\{\epsilon\sigma_0/\sigma + (\mu_0 - \mu)/\sigma\}]]\ d\epsilon < \delta/2$ for every $f_\epsilon \sim \Pi_\epsilon$. By Lemma 4, we have $\Pi_V(||v_{0Y}^{1/2} - v_Y^{1/2}||_\infty < \eta_\sigma) > 0$. Combining these results, $\Pi_{Y|\bullet}\{\sup_{X\in[A,B]} d_{KL}(f_{0Y|X}, f_{Y|X}) < \delta\} \geq \Pi_\epsilon\{d_{KL}(f_{0\epsilon_Y}, f_{\epsilon_Y}) < \delta/2\}\ \Pi_R(||r_0 - r||_\infty < \eta_\mu)\ \Pi_V(||v_{0Y}^{1/2} - v_Y^{1/2}||_\infty < \eta_\sigma) > 0$. Hence the proof of part 2 of Lemma 5.

Part 1 follows trivially from part 2 since $||r_0 - r||_\infty < \eta_\mu$ and $||v_{0Y}^{1/2} - v_Y^{1/2}||_\infty < \eta_\sigma$ imply $|r_0(X) - r(X)|_\infty < \eta_\mu$ and $|v_{0Y}^{1/2}(X) - v_Y^{1/2}(X)| < \eta_\sigma$ for any $X \in [A, B]$.

Plugging in $\psi = 0$ in Lemma 7 we get $\lim_{\tau\to 1} \int f_{0\epsilon}(\epsilon)\ \log\ [f_\epsilon(\epsilon)/\{\tau^{-1} f_\epsilon(\tau\epsilon)\}]\ d\epsilon = 0$. Using this result and arguing along the same lines, part 3 and part 4 of Lemma 5 can be proved for $m = 1$. Since $d_{KL}(f_{0\mathbf{W}_{1:m}|X}, f_{\mathbf{W}_{1:m}|X}) = \sum_{j=1}^m d_{KL}(f_{0W_j|X}, f_{W_j|X})$, the results for $m > 1$ follow trivially.

D.2.3 Proof of Theorem 1

Let $d_H(f_0, f) = [\int\{f_0^{1/2}(\mathbf{Z}) - f^{1/2}(\mathbf{Z})\}^2 d\mathbf{Z}]^{1/2}$ denote the Hellinger distance between any two densities f_0 and f. From Chapter 1 of Ghosh and Ramamoorthi (2010), we have

$$d_H^2(f_0, f) \leq ||f_0 - f||_1 \leq 2\ d_{KL}^{1/2}(f_0, f). \tag{D.1}$$

By Lemma 7 of Ghosal and van der Vaart (2007b), there exists $\lambda \in (0,1)$ such that

$$d_{KL}(f_0, f) \leq d_H^2(f_0, f)\{1 + 2\log(2/\lambda)\} + 2\int_{\{f(\mathbf{Z})<\lambda f_0(\mathbf{Z})/2\}} f_0(\mathbf{Z})\log\{f(\mathbf{Z})/f_0(\mathbf{Z})\}d\mathbf{Z} \quad \text{(D.2)}$$

For the second term in (D.2), we have

$$\int_{\{f(\mathbf{Z})<\lambda f_0(\mathbf{Z})\}} f_0(\mathbf{Z})\log\{f(\mathbf{Z})/f_0(\mathbf{Z})\}d\mathbf{Z} \leq \int_{\{f(\mathbf{Z})<\lambda f_0(\mathbf{Z})\}} f_0(\mathbf{Z})\{f(\mathbf{Z})/f_0(\mathbf{Z}) - 1\}d\mathbf{Z} \leq 0.$$

Applying (D.1) on the first term of (D.2), we have

$$d_{KL}(f_0, f) \leq \|f_0 - f\|_1 \{1 + 2\log(2/\lambda)\}. \quad \text{(D.3)}$$

Using (D.1) and (D.3), we have, for some $\lambda \in (0,1)$,

$$\begin{aligned}
&\{1 + 2\log(2/\lambda)\}^{-1} d_{KL}(f_{0Y,\mathbf{W}_{1:m}}, f_{Y,\mathbf{W}_{1:m}}) \leq ||f_{0Y,\mathbf{W}_{1:m}} - f_{Y,\mathbf{W}_{1:m}}||_1 \\
&= \int\int_{\mathbb{R}^m} |f_{0Y,\mathbf{W}_{1:m}}(Y, \mathbf{W}_{1:m}) - f_{Y,\mathbf{W}_{1:m}}(Y, \mathbf{W}_{1:m})| dY d\mathbf{W}_{1:m} \\
&\leq \int\int_{\mathbb{R}^m} \left| \int \{f_{0X}(X) - f_X(X)\}\ f_{0Y,\mathbf{W}_{1:m}|X}(Y, \mathbf{W}_{1:m}) dX \right| dY d\mathbf{W}_{1:m} \\
&+ \int\int_{\mathbb{R}^m} \left| \int f_X(X)\{f_{0,\mathbf{W}_{1:m}|X}(Y, \mathbf{W}_{1:m}) dX - f_{Y,\mathbf{W}_{1:m}|X}(Y, \mathbf{W}_{1:m})\} dX \right| dY d\mathbf{W}_{1:m} \\
&\leq \int\int_{\mathbb{R}^m} \int |f_{0X}(X) - f_X(X)|\ f_{0Y,\mathbf{W}_{1:m}|X}(Y, \mathbf{W}_{1:m}) dX dY d\mathbf{W}_{1:m} \\
&+ \int\int_{\mathbb{R}^m} \int f_X(X) \left| f_{0Y,\mathbf{W}_{1:m}|X}(Y, \mathbf{W}_{1:m}) - f_{Y,\mathbf{W}_{1:m}|X}(Y, \mathbf{W}_{1:m}) \right| dX dY d\mathbf{W}_{1:m} \\
&\leq ||f_{0X} - f_X||_1 + \sup_{X\in[A,B]} ||f_{0Y|X}(Y) f_{0\mathbf{W}_{1:m}|X}(\mathbf{W}_{1:m}) - f_{Y|X}(Y) f_{\mathbf{W}_{1:m}|X}(\mathbf{W}_{1:m})||_1 \\
&\leq ||f_{0X} - f_X||_1 + \sup_{X\in[A,B]} ||f_{0Y|X} - f_{Y|X}||_1 + \sum_{j=1}^{m} \sup_{X\in[A,B]} ||f_{0W_j|X} - f_{W_j|X}||_1 \\
&\leq 2\, d_{KL}^{1/2}(f_{0X}, f_X) + 2 \sup_{X\in[A,B]} d_{KL}^{1/2}(f_{0Y|X}, f_{Y|X}) + \sum_{j=1}^{m} \sup_{X\in[A,B]} d_{KL}^{1/2}(f_{0W_j|X}, f_{W_j|X}).
\end{aligned}$$

The conclusion of Theorem 1 now follows from Lemma 2 and part 2 and part 4 of Lemma 5.

D.3 Flexibility of the Multivariate Deconvolution Models of Section 4

This section presents analogous theoretical results showing flexibility of the multivariate deconvolution models. The statements of the results are very similar to those in Appendix D.3, but are still presented for completeness and easy reference.

Conditions 4. *1. f_0 is continuous on $\mathcal{S}$ except on a set of measure zero.*
2. The second order moments of f_0 are finite.
3. For some $r > 0$ and for all $\mathbf{z} \in \mathcal{S}$, there exist hypercubes $C_r(\mathbf{z})$ with side length r and $\mathbf{z} \in C_r(\mathbf{z})$ such that

$$\int f_0(\mathbf{z}) \; log \left\{ \frac{f_0(\mathbf{z})}{\inf_{\mathbf{t} \in C_r(\mathbf{z})} f_0(\mathbf{t})} \right\} d\mathbf{z} < \infty.$$

Let $\Pi_{\mathbf{X}}$ be a generic notation for both the MIW and the MLFA prior on the unknown density of interest based on K mixture components defined in Section 4.3. Similarly, let $\Pi_{\boldsymbol{\epsilon}}$ be a generic notation for both the MIW and the MLFA prior on the unknown density of the scaled errors defined in Section 4.4. When the measurement errors are distributed independently of $\mathbf{X}$, the support of the density of interest, say $\mathcal{X}$, may be taken to be any subset of $\mathbb{R}^p$. For conditionally heteroscedastic measurement errors, the variance functions $s_k^2(\cdot)$ that capture the conditional variability are modeled by mixtures of B-splines defined on closed intervals $[A_k, B_k]$. In this case, the support of the density of interest is assumed to be the closed hypercube $\mathcal{X} = [A_1, B_1] \times \cdots \times [A_p, B_p]$. Let $\mathcal{F}_{\mathbf{X}}$ denote the set of all densities on $\mathcal{X}$, the target class of densities to be modeled by $\Pi_{\mathbf{X}}$ and $\widetilde{\mathcal{F}}_{\mathbf{X}} \subseteq \mathcal{F}_{\mathbf{X}}$ denote the class of densities $f_{0\mathbf{X}}$ that satisfy Conditions 4. Similarly, let $\mathcal{F}_{\boldsymbol{\epsilon}}$ denote the set of all densities on $\mathbb{R}^p$ that have mean zero and $\widetilde{\mathcal{F}}_{\boldsymbol{\epsilon}} \subseteq \mathcal{F}_{\boldsymbol{\epsilon}}$ denote the class of densities $f_{0\boldsymbol{\epsilon}}$ that satisfy Conditions 4. The following lemma establishes the flexibility of the models for the density of interest and the density of the scaled measurement errors.

Lemma 8. *1. $\widetilde{\mathcal{F}}_{\mathbf{X}} \subseteq KL(\Pi_{\mathbf{X}})$ 2. $\widetilde{\mathcal{F}}_{\boldsymbol{\epsilon}} \subseteq KL(\Pi_{\boldsymbol{\epsilon}})$.*

Let $\Pi_{\mathbf{V}}$ denote the prior on the variance functions based on mixtures of B-spline basis functions defined in Section 4.4.2. The case of a univariate variance function supported on $[A, B]$ was considered in Appendix D.1. Extension to the multivariate case with variance functions supported on $\mathcal{X}$ is technically trivial.

For a given $\mathbf{X}$, let $\Pi_{\mathbf{U}|\mathbf{X}}$ denote the prior for $f_{\mathbf{U}|\mathbf{X}}$ induced by $\Pi_{\boldsymbol{\epsilon}}$ and $\Pi_{\mathbf{V}}$ under model (4.15). Define $\widetilde{\mathcal{F}}_{\mathbf{U}|\mathbf{X}} = \{f_{0\mathbf{U}|\mathbf{X}} : f_{0\mathbf{U}|\mathbf{X}}(\mathbf{U}) = \prod_{k=1}^p s_{0k}^{-1}(X_k) f_{0\boldsymbol{\epsilon}}\{\mathbf{S}_0^{-1}(\mathbf{X})\mathbf{U}\}, s_{0k}^2 \in$

$\mathcal{C}_+[A_k, B_k]$ for $k = 1, \ldots, p, f_{0\boldsymbol{\epsilon}} \in \widetilde{\mathcal{F}}_{\boldsymbol{\epsilon}}\}$. Also let $\Pi_{\mathbf{U}|\mathbf{V}}$ denote the prior for the unknown conditional density of $\mathbf{U}$ induced by $\Pi_{\boldsymbol{\epsilon}}$ and $\Pi_{\mathbf{V}}$ under model (4.15). Define $\widetilde{\mathcal{F}}_{\mathbf{U}|\bullet} = \{f_{0\mathbf{U}|\bullet} : \text{ for any given } \mathbf{X} \in \mathcal{X},\ f_{0\mathbf{U}|\bullet} = f_{0\mathbf{U}|\mathbf{X}} \in \widetilde{\mathcal{F}}_{\mathbf{U}|\mathbf{X}}\}$. Finally, let $\Pi_{\mathbf{X},\mathbf{U}}$ denote the prior for the joint density of $(\mathbf{X}, \mathbf{U})$ induced by $\Pi_{\mathbf{X}}$, $\Pi_{\boldsymbol{\epsilon}}$ and $\Pi_{\mathbf{V}}$ under model (4.15). Define $\widetilde{\mathcal{F}}_{\mathbf{X},\mathbf{U}} = \{f_{0,\mathbf{X},\mathbf{U}} : f_{0,\mathbf{X},\mathbf{U}}(\mathbf{X}, \mathbf{U}) = f_{0,\mathbf{X}}(\mathbf{X}) f_{0,\mathbf{U}|\mathbf{X}}(\mathbf{U} \mid \mathbf{X}),$ where $f_{0\mathbf{X}} \in \widetilde{\mathcal{F}}_{\mathbf{X}}$ and $f_{0\mathbf{U}|\mathbf{X}} \in \widetilde{\mathcal{F}}_{\mathbf{U}|\mathbf{X}}$ for all $\mathbf{X} \in \mathcal{X}\}$.

Lemma 9. *1.* $\widetilde{\mathcal{F}}_{\mathbf{U}|\mathbf{X}} \subseteq KL(\Pi_{\mathbf{U}|\mathbf{X}})$ *for any given* $\mathbf{X} \in \mathcal{X}$.
2. For any $f_{0\mathbf{U}|\bullet} \in \widetilde{\mathcal{F}}_{\mathbf{U}|\mathbf{V}}$, $\Pi_{\mathbf{U}|\mathbf{V}}\{\sup_{\mathbf{X}\in\mathcal{X}} d_{KL}(f_{0\mathbf{U}|\mathbf{X}}, f_{\mathbf{U}|\mathbf{X}}) < \delta\} > 0$ *for all* $\delta > 0$.
3. $\widetilde{\mathcal{F}}_{\mathbf{X},\mathbf{U}} \subseteq KL(\Pi_{\mathbf{X},\mathbf{U}})$.

Let $\Pi_{\mathbf{W}}$ denote the prior for the density of $\mathbf{W}$ induced by $\Pi_{\mathbf{X}}$, $\Pi_{\boldsymbol{\epsilon}}$ and $\Pi_{\mathbf{V}}$ under model (4.15). Also let $\widetilde{\mathcal{F}}_{\mathbf{W}} = \{f_{0\mathbf{W}} : f_{0\mathbf{W}}(\mathbf{W}) = \int f_{0\mathbf{X}}(\mathbf{X}) f_{0\mathbf{U}|\mathbf{X}}(\mathbf{W} - \mathbf{X}) d\mathbf{X}, f_{0\mathbf{X}} \in \widetilde{\mathcal{F}}_{\mathbf{X}}, f_{0\mathbf{U}|\bullet} \in \widetilde{\mathcal{F}}_{\mathbf{U}|\bullet}\}$, the class of densities $f_{0\mathbf{W}}$ that can be obtained as the convolution of two densities $f_{0\mathbf{X}}$ and $f_{0\mathbf{U}|\bullet}$, where $f_{0\mathbf{X}} \in \widetilde{\mathcal{F}}_{\mathbf{X}}$ and $f_{0\mathbf{U}|\bullet} \in \widetilde{\mathcal{F}}_{\mathbf{U}|\bullet}$. Since the supports of $\Pi_{\mathbf{X}}$ and $\Pi_{\mathbf{U}|\mathbf{V}}$ are large, it is expected that the support of $\Pi_{\mathbf{W}}$ will also be large, as is shown by the following theorem.

Theorem 2. $\widetilde{\mathcal{F}}_{\mathbf{W}} \subseteq KL(\Pi_{\mathbf{W}})$.

Proof of part 1 of Lemma 8 follows mostly from the results in Norets and Pelenis (2012), but requires modifications for the use of sophisticated priors. To accommodate the mean zero restriction on the density of the measurement errors, the proof of part 2 requires additional modifications along the lines of Pelenis (2014). We have presented these technical details in Appendix D.4 for easy reference. The proof of Lemma 9 require modifications due to multivariate set up and the use of complicated priors are also presented in Appendix D.4. The proof of Theorem 2 can be obtained by trivial modifications of the proof of Theorem 1 and is thus excluded.

We recall that our multivariate density deconvolution models are based on mixtures models with finite fixed number of components. The proofs the theoretical results pertaining to the flexibility of the multivariate deconvolution models however require that the number of mixture components K be allowed to vary over $\mathbb{N}$, the set of all positive integers, through priors, denoted by the generic $P_0(K)$, that assign positive probability to all $K \in \mathbb{N}$. Posterior computation of such methods will be computationally intensive, specially in a complicated multivariate set up like ours. In our implementation we kept the number of mixture components fixed at finite values

to reduce computational complexity. This did not necessarily come at the expense of the flexibility shown by the theoretical results of this section, since the priors $P_0(K)$ play only minor roles in the proofs. Indeed, the most important steps in the proofs of these results are actually to show that mixture models with fixed but sufficiently large number of components can approximate any target function with any desired level of accuracy provided they satisfy the above mentioned regularity conditions. With additional clauses to specify the desired degree of accuracy and lower bounds on the number of mixture components, the statements of the theoretical results presented in this section can certainly be modified to eliminate the dependence on the priors $P_0(K)$. We have not done this to keep the statements of the theoretical results short and clean and, more importantly, to be able to clearly explain the trivial roles the priors $P_0(K)$ play in the proofs that strengthen our arguments in favor of using mixtures with finite fixed number of components. These discussions were presented in Appendix C of this thesis.

D.4 Proofs of the Theoretical Results of Appendix D.3

D.4.1 Proof of Lemma 8

Proof of part 1 of Lemma 8 follows by modifications of the results of Norets and Pelenis (2012). We present here only the proof of part 2 that requires additional modifications along the lines of Pelenis (2014) to accommodate the mean zero restriction on the density of the measurement errors. The first step is to construct finite mixture models of the form

$$f_m(\mathbf{z} \mid \boldsymbol{\theta}_m) = \sum_{k=1}^{m+2} \pi_{m,k} \ \mathrm{MVN}_p(\mathbf{z} \mid \boldsymbol{\mu}_{m,k}, \boldsymbol{\Sigma}_{m,k}) \quad \text{with} \quad \sum_{k=1}^{m+2} \pi_{m,k} \boldsymbol{\mu}_{m,k} = \mathbf{0}$$

that can approximate any given density f_0 that has mean zero and satisfies Conditions 4 with any desired level of accuracy. The continuity of $f_m(\cdot \mid \boldsymbol{\theta})$ implies that the KL distance between f_0 and f_m remains small on sufficiently small open neighborhoods around $\boldsymbol{\theta}_m$. Both the MIW and the MLFA priors assign positive probability to open neighborhoods around $\boldsymbol{\theta}_m$. The conclusion of part 2 of Lemma 8 follows since the prior probability of having $(m+2)$ mixture components is also positive for all $m \in \mathbb{N}$.

Lemma 10. *For any given $f_0 \in \widetilde{\mathcal{F}}_{\boldsymbol{\epsilon}}$ and any given $\eta > 0$, there exists $\boldsymbol{\theta}_m$ such that $d_{KL}\{f_0(\cdot), f_m(\cdot|\boldsymbol{\theta}_m)\} < \eta$.*

Proof. Let $\{A_{m,k}\}_{k=1}^{m}$ be adjacent cubes with side length h_m, and $A_{m,0} = \mathbb{R}^p - \cup_{k=1}^{m} A_{m,k}$ such that $h_m \downarrow 0$ but $\cup_{k=1}^{m} A_{m,k} \uparrow \mathbb{R}^p$ as $m \to \infty$. So $\{A_{m,k}\}_{k=1}^{m}$ becomes finer but $\cup_{k=1}^{m} A_{m,k}$ covers more of $\mathbb{R}^p$. Additionally, let the partition be constructed in such a way that for all m sufficiently large, if $\boldsymbol{\epsilon} \in A_{m,0}$, then $C_r(\boldsymbol{\epsilon}) \cap A_{m,0}$ contains a hypercube $C_0(\boldsymbol{\epsilon})$ with side length $r/2$ and a vertex at $\boldsymbol{\epsilon}$; and if $\boldsymbol{\epsilon} \notin A_{m,0}$, then $C_r(\boldsymbol{\epsilon}) \cap (\mathbb{R}^p - A_{m,0})$ contains a hypercube $C_1(\boldsymbol{\epsilon})$ with side length $r/2$ and a vertex at $\boldsymbol{\epsilon}$. Consider the model

$$f_m(\mathbf{z}) \;=\; f_m(\mathbf{z} \mid \boldsymbol{\theta}_m) = \sum_{k=1}^{m+2} \pi_{m,k}\, \mathrm{MVN}_p(\mathbf{z} \mid \boldsymbol{\mu}_{m,k}, \boldsymbol{\Sigma}_{m,k}).$$

Set $\pi_{m,k} = \int_{A_{m,k}} f_0(\mathbf{z})d\mathbf{z}$ for $k = 1, 2, \ldots, m$ and $\pi_{m,k} = P_{f_0}(A_{m,0})/2 = \int_{A_{m,k}} f_0(\mathbf{z})d\mathbf{z}/2$ for $k = (m+1), (m+2)$. Then $\sum_{k=1}^{m+2} \pi_{m,k} = \int_{\mathbb{R}^p} f_0(\mathbf{z})d\mathbf{z} = 1$. Define $g(\mathbf{d}) = \sum_{k=1}^{m} \pi_{m,k}(\mathbf{c}_{m,k} + \mathbf{d}) + \int_{A_{m,0}} \mathbf{z} f_0(\mathbf{z})d\mathbf{z}$, where $\mathbf{c}_{m,k}$ is the center of $A_{m,k}$ for $k = 1, 2, \ldots, m$.

$$\begin{aligned} g(h_m \mathbf{1}_p/2) &= \sum_{k=1}^{m} \pi_{m,k}(\mathbf{c}_{m,k} + h_m \mathbf{1}_p/2) + \int_{A_{m,0}} \mathbf{z} f_0(\mathbf{z})d\mathbf{z} \\ &= \sum_{k=1}^{m} \int_{A_{m,k}} (\mathbf{c}_{m,k} + h_m \mathbf{1}_p/2) f_0(\mathbf{z})d\mathbf{z} + \int_{A_{m,0}} \mathbf{z} f_0(\mathbf{z})d\mathbf{z} \\ &\geq \sum_{k=1}^{m} \int_{A_{m,k}} \mathbf{z} f_0(\mathbf{z})d\mathbf{z} + \int_{A_{m,0}} \mathbf{z} f_0(\mathbf{z})d\mathbf{z} = \int_{\mathbb{R}^p} \mathbf{z} f_0(\mathbf{z})d\mathbf{z} = \mathbf{0}. \end{aligned}$$

Similarly $g(-h_m \mathbf{1}_p/2) \leq 0$. Since $g(\cdot)$ is continuous, there exists $\mathbf{d}_m \in [-h_m/2, h_{m/2}]^p$ such that $g(\mathbf{d}_m) = \mathbf{0}$. Set $\boldsymbol{\mu}_{m,k} = (\mathbf{c}_{m,k} + \mathbf{d}_m)$ for $k = 1, 2, \ldots, m$. Also set $\boldsymbol{\mu}_{m,m+1} = 2\int_{A_{m,0}} \mathbf{z} f_0(\mathbf{z})d\mathbf{z} / \int_{A_{m,0}} f_0(\mathbf{z})d\mathbf{z}$ and $\boldsymbol{\mu}_{m,m+2} = \mathbf{0}$ when $\int_{A_{m,0}} f_0(\mathbf{z})d\mathbf{z} > 0$, and $\boldsymbol{\mu}_{m,0} = \mathbf{0}$ otherwise. Then $\sum_{k=1}^{m+2} \pi_{m,k}\boldsymbol{\mu}_{m,k} = g(\mathbf{d}_m) = \mathbf{0}$. Also set $\boldsymbol{\Sigma}_{m,k} = \sigma_m^2 \mathrm{I}_p$ for $k = 1, 2, \ldots, m$ with $\sigma_m \to 0$, and $\Sigma_{m,m+1} = \Sigma_{m,m+2} = \sigma_0^2 \mathrm{I}_p$.

Consider a sequence $\{\delta_m\}_{m=1}^{\infty}$ satisfying $\delta_m > 6p^{1/2}h_m$ and $\delta_m \to 0$. Fix $\boldsymbol{\epsilon} \in \mathbb{R}^p$. Define $C_{\delta_m}(\boldsymbol{\epsilon}) = [\boldsymbol{\epsilon} - \delta_m \mathbf{1}_p/2, \boldsymbol{\epsilon} + \delta_m \mathbf{1}_p/2]$. For m sufficiently large $C_{\delta_m}(\boldsymbol{\epsilon}) \subseteq \cup_{k=1}^{m} A_{m,k}$, $C_{\delta_m}(\boldsymbol{\epsilon}) \cap A_{m,0} = \phi$ and the set $\{k : 1 \leq k \leq m, A_{m,k} \subset C_{\delta_m}(\boldsymbol{\epsilon})\}$ is non-empty. For

$k = 1, \ldots, m$, when $A_{m,k} \subset C_{\delta_m}(\boldsymbol{\epsilon})$, $\pi_{m,k} \geq \inf_{\mathbf{z} \in C_{\delta_m}(\boldsymbol{\epsilon})} f_0(\mathbf{z}) h_m^p$. Therefore,

$$\begin{aligned} f_m(\boldsymbol{\epsilon}) &\geq \sum_{\{k:1 \leq k \leq m, A_{m,k} \subset C_{\delta_m}(\boldsymbol{\epsilon})\}} \pi_{m,k} \ \mathrm{MVN}_p(\boldsymbol{\epsilon} \mid \boldsymbol{\mu}_{m,k}, \sigma_m^2 \mathrm{I}_p) \\ &\geq \inf_{z \in C_{\delta_m}(\boldsymbol{\epsilon})} f_0(\mathbf{z}) \sum_{\{k:A_{m,k} \subset C_{\delta_m}(\boldsymbol{\epsilon})\}} h_m^p \ \mathrm{MVN}_p(\boldsymbol{\epsilon} \mid \mathbf{c}_{m,k} + \mathbf{d}_m, \sigma_m^2 \mathrm{I}_p) \\ &\geq \inf_{z \in C_{\delta_m}(\boldsymbol{\epsilon})} f_0(\mathbf{z}) \quad \left\{ 1 - \frac{6p^{3/2} h_m \delta_m^{p-1}}{(2\pi)^{p/2} \sigma_m^p} - \frac{8p\sigma_m}{(2\pi)^{1/2} \delta_m} \right\}, \end{aligned}$$

where the last step follows from Lemma 1 and Lemma 2 of Norets and Pelenis (2012). Let h_m, δ_m, σ_m further satisfy $h_m/\sigma_m^p \to 0, \sigma_m/\delta_m \to 0$. Then for any $\eta > 0$ there exists an M_1 large enough such that for all $m > M_1$

$$f_m(\boldsymbol{\epsilon}) \geq \inf_{\mathbf{z} \in C_{\delta_m}(\boldsymbol{\epsilon})} f_0(\mathbf{z}) \cdot (1 - \eta).$$

Without loss of generality, we may assume $f_0(\boldsymbol{\epsilon}) > 0$. Since $f_0(\cdot)$ is continuous and $\delta_m \to 0$, there also exists an M_2 such that for all $m > M_2$ we have $\inf_{\mathbf{z} \in C_{\delta_m}(\boldsymbol{\epsilon})} f_0(\mathbf{z}) > 0$ and

$$\frac{f_0(\boldsymbol{\epsilon})}{\inf_{\mathbf{z} \in C_{\delta_m}(\boldsymbol{\epsilon})} f_0(\mathbf{z})} \leq (1 + \eta).$$

Therefore, for all $m > \max\{M_1, M_2\}$, we have

$$1 \leq \max\left\{1, \frac{f_0(\boldsymbol{\epsilon})}{f_m(\boldsymbol{\epsilon})}\right\} \leq \max\left\{1, \frac{f_0(\boldsymbol{\epsilon})}{\inf_{z \in C_{\delta_m}(\boldsymbol{\epsilon})} f_0(z) \cdot (1 - \eta)}\right\} \leq \frac{(1 + \eta)}{(1 - \eta)}.$$

Thus, $\log \max\{1, f_0(\boldsymbol{\epsilon})/f_m(\boldsymbol{\epsilon})\} \to 0$ as $m \to \infty$. Pointwise convergence is thus established. Next, we will find an integrable upper bound for $\log \max\{1, f_0(\boldsymbol{\epsilon})/f_m(\boldsymbol{\epsilon})\}$.

For point wise convergence we can assume $\boldsymbol{\epsilon} \notin A_{m,0}$ for sufficiently large m. But to find integrable upper bound, we have to consider both the cases $\boldsymbol{\epsilon} \in A_{m,0}$ and $\boldsymbol{\epsilon} \notin A_{m,0}$. When $\boldsymbol{\epsilon} \in A_{m,0}$, we have $\mathrm{P}_{f_0}(A_{m,0}) = \int_{A_{m,0}} f_0(\mathbf{z}) d\mathbf{z} \geq \int_{A_{m,0} \cap C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z}) d\mathbf{z} \geq \lambda\{A_{m,0} \cap C_r(\boldsymbol{\epsilon})\} \inf_{\mathbf{z} \in A_{m,0} \cap C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z}) \geq (r/2)^p \inf_{\mathbf{z} \in C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z})$, since $\lambda\{A_{m,0} \cap C_r(\boldsymbol{\epsilon})\} \geq \lambda\{C_0(\boldsymbol{\epsilon})\} \geq (r/2)^p$. Using part 4 of Conditions 4 and Lemma 1 and Lemma 2 of

Norets and Pelenis (2012) again, if $\boldsymbol{\epsilon} \notin A_{m,0}$, for m sufficiently large

$$\begin{aligned}
&\sum_{\{k:A_{m,k}\subset C_r(\boldsymbol{\epsilon})\}} h_m^p \ \mathrm{MVN}_p(\boldsymbol{\epsilon} \mid \boldsymbol{\mu}_{m,k}, \sigma_m^2 \mathrm{I}_p) \geq \sum_{\{k:A_{m,k}\subset C_1(\boldsymbol{\epsilon})\}} h_m^p \ \mathrm{MVN}_p(\boldsymbol{\epsilon} \mid \boldsymbol{\mu}_{m,k}, \sigma_m^2 \mathrm{I}_p) \\
&\quad \geq \int_{C_1(\boldsymbol{\epsilon})} \mathrm{MVN}_p(\mathbf{z} \mid \boldsymbol{\epsilon}, \sigma_m^2 \mathrm{I}_p) d\mathbf{z} - \frac{3p^{3/2}(r/2)^{p-1} h_m}{(2\pi)^{p/2}\sigma_m^p} \\
&\quad \geq \left\{ \frac{1}{2^p} - \frac{8p\sigma_m}{2^p(2\pi)^{1/2} r} - \frac{3p^{3/2} h_m r^{p-1}}{2^{p-1}(2\pi)^{p/2}\sigma_m^p} \right\} \geq \frac{1}{2^{p+1}},
\end{aligned}$$

This implies

$$\begin{aligned}
f_m(\boldsymbol{\epsilon}) &= \sum_{k=1}^{m} P_{f_0}(A_{m,k}) \ \mathrm{MVN}_p(\boldsymbol{\epsilon} \mid \boldsymbol{\mu}_{m,k}, \sigma_m^2 \mathrm{I}_p) \\
&\qquad + \sum_{k=m+1}^{m+2} (1/2) P_{f_0}(A_{m,0}) \ \mathrm{MVN}_p(\boldsymbol{\epsilon} \mid \boldsymbol{\mu}_{m,k}, \sigma_0^2 \mathrm{I}_p) \\
&\geq \sum_{k=1}^{m} P_{f_0}(A_{m,k}) \ \mathrm{MVN}_p(\boldsymbol{\epsilon} \mid \boldsymbol{\mu}_{m,k}, \sigma_m^2 \mathrm{I}_p) + (1/2) P_{f_0}(A_{m,0}) \ \mathrm{MVN}_p(\boldsymbol{\epsilon} \mid \mathbf{0}, \sigma_0^2 \mathrm{I}_p) \\
&\geq \{1 - 1(\boldsymbol{\epsilon} \in A_{m,0})\} \inf_{\mathbf{z}\in C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z}) \sum_{\{k:A_{m,k}\subset C_r(\boldsymbol{\epsilon})\}} \lambda(A_{m,k}) \ \mathrm{MVN}_p(\boldsymbol{\epsilon} \mid \boldsymbol{\mu}_{m,k}, \sigma_m^2 \mathrm{I}_p) \\
&\quad + 1(\boldsymbol{\epsilon} \in A_{m,0}) (1/2) P_{f_0}(A_{m,0}) \ \mathrm{MVN}_p(\boldsymbol{\epsilon} \mid \mathbf{0}, \sigma_0^2 \mathrm{I}_p) \\
&\geq (1/2)\{1 - 1(\boldsymbol{\epsilon} \in A_{m,0})\} \inf_{\mathbf{z}\in C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z}) \\
&\quad + 1(\boldsymbol{\epsilon} \in A_{m,0}) \ (1/2)(r/2)^p \ \mathrm{MVN}_p(\boldsymbol{\epsilon} \mid \mathbf{0}, \sigma_0^2 \mathrm{I}_p) \inf_{\mathbf{z}\in C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z}) \\
&\geq (1/2)(r/2)^p \ \mathrm{MVN}_p(\boldsymbol{\epsilon} \mid \mathbf{0}, \sigma_0^2 \mathrm{I}_p) \inf_{\mathbf{z}\in C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z}).
\end{aligned}$$

The last step followed by choosing σ_0^2 large enough so that $(r/2)^p \sup_{\boldsymbol{\epsilon}\in\mathbb{R}^p} \mathrm{MVN}_p(\boldsymbol{\epsilon} \mid \mathbf{0}, \sigma_0^2 \mathrm{I}_p) < (r/2)^p \ \sigma_0^{-p} < 2^{-(p+1)} < 1$. Therefore,

$$\begin{aligned}
&\log \max\left\{1, \frac{f_0(\boldsymbol{\epsilon})}{f_m(\boldsymbol{\epsilon})}\right\} \leq \log \max\left\{1, \frac{f_0(\boldsymbol{\epsilon})}{(1/2)(r/2)^p \ \mathrm{MVN}_p(\boldsymbol{\epsilon} \mid \mathbf{0}, \sigma_0^2 \mathrm{I}_p) \ \inf_{\mathbf{z}\in C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z})}\right\} \\
&\leq -\log\left\{(1/2)(r/2)^p \ \mathrm{MVN}_p(\boldsymbol{\epsilon} \mid \mathbf{0}, \sigma_0^2 \mathrm{I}_p)\right\} + \log\left\{\frac{f_0(\boldsymbol{\epsilon})}{\inf_{\mathbf{z}\in C_r(\boldsymbol{\epsilon})} f_0(\mathbf{z})}\right\}.
\end{aligned}$$

The first and the second terms are integrable by part 2 and part 3 of Conditions 4, respectively. Since $\int f_0(\boldsymbol{\epsilon}) \log\{f_{0\boldsymbol{\epsilon}}/f_m(\boldsymbol{\epsilon})\} d\boldsymbol{\epsilon} \leq \int f_0(\boldsymbol{\epsilon}) \log \max\{1, f_{0\boldsymbol{\epsilon}}/f_m(\boldsymbol{\epsilon})\} d\boldsymbol{\epsilon}$, the proof of Lemma 10 is completed applying dominated convergence theorem (DCT).

Let $\eta > 0$ be given. By Lemma 10, there exists $\boldsymbol{\theta}_m^\star = (\boldsymbol{\pi}_{1:(m+2)}^\star, \boldsymbol{\mu}_{1:(m+2)}^\star, \boldsymbol{\Sigma}_{1:(m+2)}^\star)$ with $\boldsymbol{\Sigma}_k^\star = \sigma_m^{2\star}\mathrm{I}_p$ for $k = 1, \ldots, m$ and $\boldsymbol{\Sigma}_k^\star = \sigma_0^{2\star}\mathrm{I}_p$ for $k = (m+1), (m+2)$ such that $d_{KL}\{f_0(\cdot), f_m(\cdot \mid \boldsymbol{\theta}_m^\star)\} < \eta/2$. We have, for any $\boldsymbol{\theta}_m$,

$$\begin{aligned}
&\int f_0(\boldsymbol{\epsilon}) \log\left\{\frac{f_0(\boldsymbol{\epsilon})}{f_m(\boldsymbol{\epsilon} \mid \boldsymbol{\theta}_m)}\right\} d\boldsymbol{\epsilon} \\
&\quad = \int f_0(\boldsymbol{\epsilon}) \log\left\{\frac{f_0(\boldsymbol{\epsilon})}{f_m(\boldsymbol{\epsilon} \mid \boldsymbol{\theta}_m^\star)}\right\} d\boldsymbol{\epsilon} + \int f_0(\boldsymbol{\epsilon}) \log\left\{\frac{f_m(\boldsymbol{\epsilon} \mid \boldsymbol{\theta}_m^\star)}{f_m(\boldsymbol{\epsilon} \mid \boldsymbol{\theta}_m)}\right\} d\boldsymbol{\epsilon}.
\end{aligned}$$

Let the second term in the above expression be denoted by $g(\boldsymbol{\theta}_m)$. The priors puts positive mass on arbitrarily small open neighborhoods around $\boldsymbol{\theta}_m^\star$. The result will follow if there exists an open neighborhood $\mathcal{N}(\boldsymbol{\theta}_m^\star)$ around $\boldsymbol{\theta}_m^\star$ such that $\sup_{\boldsymbol{\theta}_m \in \mathcal{N}(\boldsymbol{\theta}_m^\star)} g(\boldsymbol{\theta}_m) < \eta/2$. Since $g(\boldsymbol{\theta}_m^\star) = 0$, it suffices to show that the function $g(\boldsymbol{\theta}_m)$ is continuous at $\boldsymbol{\theta}_m^\star$. Now $g(\boldsymbol{\theta})$ is continuous at $\boldsymbol{\theta}_m^\star$ if for every sequence $\{\boldsymbol{\theta}_{m,n}\}_{n=1}^\infty$ with $\boldsymbol{\theta}_{m,n} \to \boldsymbol{\theta}_m^\star$, we have $g(\boldsymbol{\theta}_{m,n}) \to g(\boldsymbol{\theta}_m^\star)$. For all $\boldsymbol{\epsilon} \in \mathbb{R}^p$, we have $\log\{f_m(\boldsymbol{\epsilon} \mid \boldsymbol{\theta}_{m,n}^\star)/f_m(\boldsymbol{\epsilon} \mid \boldsymbol{\theta}_m)\} \to 0$ as $\boldsymbol{\theta}_{m,n} \to \boldsymbol{\theta}_m^\star$. Continuity of $g(\boldsymbol{\theta}_m)$ at $\boldsymbol{\theta}_m^\star$ will follow from DCT if we can show that $|f_m(\boldsymbol{\epsilon} \mid \boldsymbol{\theta}_m^\star)/f_m(\boldsymbol{\epsilon} \mid \boldsymbol{\theta}_{m,n})|$ has an integrable with respect to f_0 upper bound.

Since $\boldsymbol{\theta}_{m,n} \to \boldsymbol{\theta}_m^\star$, for any open neighborhood $\mathcal{N}(\boldsymbol{\theta}_m^\star)$ around $\boldsymbol{\theta}_m^\star$, we must have $\boldsymbol{\theta}_{m,n} \in \mathcal{N}(\boldsymbol{\theta}_m^\star)$ for all n sufficiently large. Let $\boldsymbol{\theta}_m = (\boldsymbol{\pi}_{1:(m+2)}, \boldsymbol{\mu}_{1:(m+2)}, \boldsymbol{\Sigma}_{1:(m+2)}) \in \mathcal{N}(\boldsymbol{\theta}_m^\star)$. Since the eigenvalues of a real symmetric matrix depend continuously on the matrix, we must have $(\lambda_1(\boldsymbol{\Sigma}_k), \lambda_p(\boldsymbol{\Sigma}_k)) \subset (\underline{\sigma}_m^{2\star}, \overline{\sigma}_m^{2\star})$ for $k = 1, \ldots, m$ and $(\lambda_1(\boldsymbol{\Sigma}_k), \lambda_p(\boldsymbol{\Sigma}_k)) \subset (\underline{\sigma}_0^{2\star}, \overline{\sigma}_0^{2\star})$ for $k = (m+1), (m+2)$, where $\underline{\sigma}_m^{2\star} < \sigma_m^{2\star} < \overline{\sigma}_m^{2\star}$ and $\underline{\sigma}_0^{2\star} < \sigma_0^{2\star} < \overline{\sigma}_0^{2\star}$. Let $\underline{\sigma}^{2\star} = \min\{\underline{\sigma}_m^{2\star}, \underline{\sigma}_0^{2\star}\}$ and $\overline{\sigma}^{2\star} = \max\{\overline{\sigma}_m^{2\star}, \overline{\sigma}_0^{2\star}\}$. Then $(\lambda_1(\boldsymbol{\Sigma}_k), \lambda_p(\boldsymbol{\Sigma}_k)) \subset (\underline{\sigma}^{2\star}, \overline{\sigma}^{2\star})$ for $k = 1, \ldots, (m+2)$. Similarly, for some finite $\mu^\star$, we must have $\boldsymbol{\mu}_{m,k} \in (-\mu^\star \mathbf{1}_p, \mu^\star \mathbf{1}_p) = \mathcal{N}_{\mu^\star}$ for $k = 1, \ldots, (m+2)$. For any real positive definite matrix $\boldsymbol{\Sigma}$, we have $\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z} \leq \lambda_1^{-1}(\boldsymbol{\Sigma})\,\|\mathbf{z}\|^2$. Therefore, for any $\boldsymbol{\epsilon} \in \mathbb{R}^p$ and for all $k = 1, \ldots, (m+2)$, we must have $(\boldsymbol{\epsilon} - \boldsymbol{\mu}_{m,k})^{\mathrm{T}}\boldsymbol{\Sigma}_{m,k}^{-1}(\boldsymbol{\epsilon} - \boldsymbol{\mu}_{m,k}) \leq \underline{\sigma}^{-2\star}\{1(\boldsymbol{\epsilon} \in \mathcal{N}_{\mu^\star})2^p\mu^{\star p} + 1(\boldsymbol{\epsilon} \notin \mathcal{N}_{\mu^\star})\,\|\boldsymbol{\epsilon} + \mathrm{sign}(\boldsymbol{\epsilon})\mu^\star\|^2\}$, where $\mathrm{sign}(\boldsymbol{\epsilon}) = \{\mathrm{sign}(\epsilon_1), \ldots, \mathrm{sign}(\epsilon_p)\}^{\mathrm{T}}$. Therefore, for any $\boldsymbol{\theta}_m \in \mathcal{N}(\boldsymbol{\theta}_m^\star)$, we have

$$\begin{aligned}
[1(\boldsymbol{\epsilon} \in \mathcal{N}_{\mu^\star})\mathrm{MVN}_p(2\mu^\star \mathbf{1}_p \mid \mathbf{0}, \underline{\sigma}^{2\star}\mathrm{I}_p) + 1(\boldsymbol{\epsilon} \notin \mathcal{N}_{\mu^\star})\mathrm{MVN}_p\{\boldsymbol{\epsilon} + \mathrm{sign}(\boldsymbol{\epsilon})\mu^\star \mid \mathbf{0}, \underline{\sigma}^{2\star}\mathrm{I}_p\}]/\overline{\sigma}^\star \\
\leq f_m(\boldsymbol{\epsilon} \mid \boldsymbol{\theta}_m) \leq 1/\underline{\sigma}^\star.
\end{aligned}$$

The upper bound is a constant and the logarithm of the lower bound is integrable

since, by part 2 of Conditions 4, the second order moments of $\boldsymbol{\epsilon}$ exist. An f_0 integrable upper bound for the function $\sup_{\boldsymbol{\theta}_m \in \mathcal{N}(\boldsymbol{\theta}_m^\star)} |f_m(\boldsymbol{\epsilon} \mid \boldsymbol{\theta}_m)|$ thus exists. Finally, DCT applies because

$$\begin{aligned}\int f_0(\boldsymbol{\epsilon}) \ \left|\log\left\{\frac{f_m(\boldsymbol{\epsilon} \mid \boldsymbol{\theta}_m^\star)}{f_m(\boldsymbol{\epsilon} \mid \boldsymbol{\theta}_{m,n})}\right\}\right| d\boldsymbol{\epsilon} &\leq \sup_{\boldsymbol{\theta}_m \in \mathcal{N}(\boldsymbol{\theta}_m^\star)} \int f_0(\boldsymbol{\epsilon}) \ \left|\log\left\{\frac{f_m(\boldsymbol{\epsilon} \mid \boldsymbol{\theta}_m^\star)}{f_m(\boldsymbol{\epsilon} \mid \boldsymbol{\theta}_m)}\right\}\right| d\boldsymbol{\epsilon} \\ &\leq 2 \sup_{\boldsymbol{\theta}_m \in \mathcal{N}(\boldsymbol{\theta}_m^\star)} \int f_0(\boldsymbol{\epsilon}) \ |f_m(\boldsymbol{\epsilon} \mid \boldsymbol{\theta}_m)| \, d\boldsymbol{\epsilon}.\end{aligned}$$

The conclusion of part 2 of Lemma 8 follows since the prior probability of having $(m+2)$ mixture components is positive for all $m \in \mathbb{N}$.

Let $P_{\boldsymbol{\epsilon},K}\{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mid \boldsymbol{\pi}_{1:K}, \boldsymbol{\mu}_{1:K}, \boldsymbol{\Sigma}_{1:K}\} = \sum_{k=1}^K \pi_k \delta_{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\delta_{\boldsymbol{\theta}}$ denotes a point mass at $\boldsymbol{\theta}$. Keeping the hyper-parameters implicit, $P_0(\boldsymbol{\pi}_{1:K}, \boldsymbol{\mu}_{1:K}, \boldsymbol{\Sigma}_{1:K}) = P_{0\pi}(\boldsymbol{\pi}_{1:K}) P_{0\mu}(\boldsymbol{\mu}_{1:K} \mid \boldsymbol{\pi}_{1:K}) P_{0\Sigma}(\boldsymbol{\Sigma}_{1:K})$. Denoting $P_{\boldsymbol{\epsilon},K}\{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mid \boldsymbol{\pi}_{1:K}, \boldsymbol{\mu}_{1:K}, \boldsymbol{\Sigma}_{1:K}\}$ simply by $P_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let c be a generic for constants that are not of direct interest. For any square matrix $\mathbf{A}$ of order p, let $\lambda_1(\mathbf{A}) \leq \cdots \leq \lambda_p(\mathbf{A})$ denote the ordered eigenvalues of $\mathbf{A}$. The following lemma proves some properties of $P_{\boldsymbol{\epsilon},K}$ and $f_{\boldsymbol{\epsilon}}$.

Lemma 11. *1.* $\int \|\boldsymbol{\mu}\|_2^2 \, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) < \infty$ *a.s.* *2.* $\int \lambda_1^{-1}(\boldsymbol{\Sigma}) dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) < \infty$ *a.s.* *3.* $\int |\boldsymbol{\Sigma}|^{-1/2} \, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) < \infty$ *a.s.*

Proof. 1. The prior $P_{0\mu}(\boldsymbol{\mu}_{1:K} \mid \boldsymbol{\pi}_{1:K})$ is of the form (4.14), that is, $P_{0\mu}(\boldsymbol{\mu}_{1:K} \mid \boldsymbol{\pi}_{1:K}) = \text{MVN}_{Kp}(\mathbf{0}, \boldsymbol{\Sigma}^0 - \boldsymbol{\Sigma}^0_{1,R} \boldsymbol{\Sigma}^{-1}_{R,R} \boldsymbol{\Sigma}^0_{R,1})$, where $\boldsymbol{\Sigma}^0$ is a $Kp \times Kp$ block-diagonal matrix independent of $\boldsymbol{\pi}_{1:K}$, all k principal blocks of order $p \times p$ being $\boldsymbol{\Sigma}_0$. The matrix $\boldsymbol{\Sigma}^0_{1,R} \boldsymbol{\Sigma}^{-1}_{R,R} \boldsymbol{\Sigma}^0_{R,1}$ depends on $\boldsymbol{\pi}_{1:K}$ and is nonnegative definite so that its diagonal elements are all nonnegative. Let $\boldsymbol{\Sigma}_0 = ((\sigma_{0,ij}))$ and $\boldsymbol{\Sigma}^0_{1,R} \boldsymbol{\Sigma}^{-1}_{R,R} \boldsymbol{\Sigma}^0_{R,1} = ((\sigma_{R,ij}))$. Then, $\int \|\boldsymbol{\mu}_k\|_2^2 \, dP_{0\mu}(\boldsymbol{\mu}_{1:K} \mid \boldsymbol{\pi}_{1:K}) = \left\{\sum_{j=1}^p \sigma_{0,jj} - \sum_{j=(k-1)p+1}^{kp} \sigma_{R,jj}\right\} \leq \sum_{j=1}^p \sigma_{0,jj} = \text{trace}(\boldsymbol{\Sigma}_0)$. Therefore,

$$\begin{aligned}&\int\int \|\boldsymbol{\mu}\|_2^2 \, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) dP_0(\boldsymbol{\pi}_{1:K}, \boldsymbol{\mu}_{1:K}, \boldsymbol{\Sigma}_{1:K}) \\ &\quad = \sum_{k=1}^K \int \pi_k \|\boldsymbol{\mu}_k\|_2^2 \, dP_{0\mu}(\boldsymbol{\mu}_{1:K} \mid \boldsymbol{\pi}_{1:K}) dP_{0\pi}(\boldsymbol{\pi}_{1:K}) \leq \text{trace}(\boldsymbol{\Sigma}_0) < \infty.\end{aligned}$$

2. We have $\int\int \lambda_1^{-1}(\boldsymbol{\Sigma}) dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) dP_0(\boldsymbol{\pi}_{1:K}, \boldsymbol{\mu}_{1:K}, \boldsymbol{\Sigma}_{1:K}) = \int \lambda_1^{-1}(\boldsymbol{\Sigma}) dP_{0\Sigma}(\boldsymbol{\Sigma})$.

When $\boldsymbol{\Sigma} \sim \mathrm{IW}_p(\nu_0, \boldsymbol{\Psi}_0)$, $\boldsymbol{\Psi}_0^{-1/2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}_0^{-1/2} \sim \mathrm{W}_p(\nu_0, \mathrm{I})$ and $\mathrm{trace}(\boldsymbol{\Psi}_0^{-1}\boldsymbol{\Sigma}^{-1}) = \mathrm{trace}(\boldsymbol{\Psi}_0^{-1/2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}_0^{-1/2}) \sim \chi^2_{p\nu_0}$. Here $\mathrm{W}_p(\nu, \boldsymbol{\Psi})$ denotes a Wishart distribution with degrees of freedom ν and mean $\nu\boldsymbol{\Psi}$. For any two positive semidefinite matrices $\mathbf{A}$ and $\mathbf{B}$, we have $\lambda_1(\mathbf{A})\mathrm{trace}(\mathbf{B}) \leq \mathrm{trace}(\mathbf{AB}) \leq \lambda_p(\mathbf{A})\mathrm{trace}(\mathbf{B})$. Therefore, $\lambda_1(\boldsymbol{\Psi}_0^{-1})E\{\mathrm{trace}(\boldsymbol{\Sigma}^{-1})\} \leq E\{\mathrm{trace}(\boldsymbol{\Psi}_0^{-1}\boldsymbol{\Sigma}^{-1})\} = p\nu_0$. Hence, $\int \lambda_1^{-1}(\boldsymbol{\Sigma})dP_{0\Sigma}(\boldsymbol{\Sigma}) = E\lambda_p(\boldsymbol{\Sigma}^{-1}) \leq E\{\mathrm{trace}(\boldsymbol{\Sigma}^{-1})\} < \infty$.

When $\boldsymbol{\Sigma} = (\boldsymbol{\Omega} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}})$ with $\boldsymbol{\Omega} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, we have $\mathrm{trace}(\boldsymbol{\Sigma}^{-1}) = \mathrm{trace}\{\boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}(\mathrm{I}_p + \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\} \leq \mathrm{trace}(\boldsymbol{\Omega}^{-1}) = \sum_{j=1}^p \sigma_j^{-2}$, where $\boldsymbol{\Gamma}$ is a $p \times p$ matrix satisfying $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathrm{T}} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$. Thus, $\int \lambda_1^{-1}(\boldsymbol{\Sigma})dP_{0\Sigma}(\boldsymbol{\Sigma}_{1:K}) = E\lambda_p(\boldsymbol{\Sigma}^{-1}) \leq E\{\mathrm{trace}(\boldsymbol{\Sigma}^{-1})\} \leq \sum_{j=1}^p E\sigma_j^{-2} < \infty$ whenever $\sigma_j^2 \sim \text{Inv-Ga}(a, b)$ with $a > 1$.

3. When $\boldsymbol{\Sigma} \sim \mathrm{IW}_p(\nu_0, \boldsymbol{\Psi}_0)$, $\lambda_1^{p/2}(\boldsymbol{\Psi}_0^{-1})E\{\mathrm{trace}(\boldsymbol{\Sigma}^{-1})\}^{p/2} \leq E\{\mathrm{trace}(\boldsymbol{\Psi}_0^{-1}\boldsymbol{\Sigma}^{-1})\}^{p/2} < \infty$. Hence, $\int |\boldsymbol{\Sigma}|^{-1/2} dP_{0\Sigma}(\boldsymbol{\Sigma}) = \int \prod_{j=1}^p \lambda_j^{1/2}(\boldsymbol{\Sigma}^{-1})dP_{0\Sigma}(\boldsymbol{\Sigma}) \leq \int \lambda_p^{p/2}(\boldsymbol{\Sigma}^{-1})dP_{0\Sigma}(\boldsymbol{\Sigma}) = E\lambda_p^{p/2}(\boldsymbol{\Sigma}^{-1}) \leq E\{\mathrm{trace}(\boldsymbol{\Sigma}^{-1})\}^{p/2} < \infty$.

For any two positive semidefinite matrix $\mathbf{A}$ and $\mathbf{B}$, we have $|\mathbf{A} + \mathbf{B}| \geq |\mathbf{A}|$. Therefore, when $\boldsymbol{\Sigma} = (\boldsymbol{\Omega} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}})$, we have $\int |\boldsymbol{\Sigma}|^{-1/2} dP_{0\Sigma}(\boldsymbol{\Sigma}_{1:K}) \leq \int |\boldsymbol{\Omega}|^{-1/2} dP_{0\Sigma}(\boldsymbol{\Sigma}_{1:K}) = \int \prod_{j=1}^p \sigma_j^{-1} dP_{0\Sigma}(\boldsymbol{\Sigma}_{1:K}) = \prod_{j=1}^p E\sigma_j^{-1} < \infty$, whenever $\sigma_j^2 \sim \text{Inv-Ga}(a, b)$ independently.

The following lemma is a multivariate analogue of Lemma 7 and proves a property of $f_{\boldsymbol{\epsilon}} = \int\int f_{c\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma})dP_0(K)$. Here $P_0(K)$ denotes the prior on K, the number of mixture components. The proof follows by minor modifications of the proof of Lemma 7.

Lemma 12. *Let $f_{0\boldsymbol{\epsilon}} \in \widetilde{\mathcal{F}}_{\boldsymbol{\epsilon}}$ and $f_{\boldsymbol{\epsilon}} \sim \Pi_{\boldsymbol{\epsilon}}$ and $\mathbf{D}(\boldsymbol{\tau}) = diag(\tau_1, \tau_2, \ldots, \tau_p)$. Then*

$$\lim_{\boldsymbol{\tau} \to \mathbf{1}} \int f_{0\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) \; log\left[\frac{f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})}{|\mathbf{D}(\boldsymbol{\tau})|^{-1} f_{\boldsymbol{\epsilon}}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}\}}\right] d\boldsymbol{\epsilon} = 0.$$

Proof. We have $|\mathbf{D}(\boldsymbol{\tau})|^{-1} f_{c\boldsymbol{\epsilon}}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}\} \to f_{c\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$ as $\boldsymbol{\tau} \to \mathbf{1}$. Since $\boldsymbol{\tau} \to \mathbf{1}$, without loss of generality, we may assume $|\mathbf{D}(\boldsymbol{\tau})| > 1/2$. Define $c = \int |\boldsymbol{\Sigma}|^{-1/2} dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) < \infty$. Also $\int |\mathbf{D}(\boldsymbol{\tau})|^{-1} f_{c\boldsymbol{\epsilon}}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon} \mid \boldsymbol{\theta}\}dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \leq \int 2(2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) < 2c < \infty$. Applying DCT, $|\mathbf{D}(\boldsymbol{\tau})|^{-1} f_{\boldsymbol{\epsilon}}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}\} \to f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$ as $\boldsymbol{\tau} \to \mathbf{1}$. Therefore, for any $\boldsymbol{\epsilon} \in \mathbb{R}$,

$$\log\left[\frac{f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})}{|\mathbf{D}(\boldsymbol{\tau})|^{-1} f_{\boldsymbol{\epsilon}}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}\}}\right] \to 0 \quad \text{as } \boldsymbol{\tau} \to \mathbf{1}.$$

To find an integrable with respect to $f_{0\boldsymbol{\epsilon}}$ upper bound for $\log\left[|\mathbf{D}(\boldsymbol{\tau})|\, f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})/f_{\boldsymbol{\epsilon}}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}\}\right]$, we use Lemma 11. To do so, we can ignore the prior $P_0(K)$ since the upper bounds obtained in Lemma 11 do not depend on the specific choice of K. We have, using part 3 of Lemma 11,

$$\int |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}-\boldsymbol{\mu}\}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}-\boldsymbol{\mu}\}\right] dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\ \leq \int |\boldsymbol{\Sigma}|^{-1/2}\, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \leq c.$$

Since $\boldsymbol{\tau} \to \mathbf{1}$, without loss of generality we may also assume $\tau_k < 2$ for all k. Therefore,

$$\begin{aligned}
|\log\, f_{\boldsymbol{\epsilon}}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}\}| &\leq \log(2\pi)^{p/2} \\
&\quad + \left|\log \int |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}-\boldsymbol{\mu}\}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}-\boldsymbol{\mu}\}\right] dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma})\right| \\
&\leq \log(2\pi)^{p/2} + |\log\, c| \\
&\quad - \log \int c^{-1}\, |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}-\boldsymbol{\mu}\}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}-\boldsymbol{\mu}\}\right] dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
&\leq \log\{c(2\pi)^{p/2}\} + |\log\, c| + \frac{1}{2}\int \log|\boldsymbol{\Sigma}|\, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
&\quad + \frac{1}{2}\int \{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}-\boldsymbol{\mu}\}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\{\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}-\boldsymbol{\mu}\} dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
&\leq \log\{c(2\pi)^{p/2}\} + |\log\, c| + \frac{1}{2}\int \log|\boldsymbol{\Sigma}|\, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
&\quad + \frac{1}{2}\int \|\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}-\boldsymbol{\mu}\|_2^2\, \lambda_1^{-1}(\boldsymbol{\Sigma}) dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
&\leq \log\{c(2\pi)^{p/2}\} + |\log\, c| + \frac{1}{2}\int \log|\boldsymbol{\Sigma}|\, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
&\quad + \int \{\|\mathbf{D}(\boldsymbol{\tau})\boldsymbol{\epsilon}\|_2^2 + \|\boldsymbol{\mu}\|_2^2\}\lambda_1^{-1}(\boldsymbol{\Sigma}) dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
&\leq \log\{c(2\pi)^{p/2}\} + |\log\, c| + \frac{1}{2}\int \log|\boldsymbol{\Sigma}|\, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
&\quad + \|2\boldsymbol{\epsilon}\|_2^2 \int \lambda_1^{-1}(\boldsymbol{\Sigma}) dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma}) + \int \|\boldsymbol{\mu}\|_2^2\, dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \int \lambda_1^{-1}(\boldsymbol{\Sigma}) dP_{\boldsymbol{\epsilon},K}(\boldsymbol{\mu},\boldsymbol{\Sigma}),
\end{aligned}$$

where the third step followed from application of Jensen's inequality on $g(Z) = -\log Z$. The regularity assumptions on $f_{0\boldsymbol{\epsilon}}$ and Lemma 11 imply that the RHS above is $f_{0\boldsymbol{\epsilon}}$ integrable. An application of DCT again completes the proof.

Let $f_{\mathbf{U}|\mathbf{S}}$ denote the density of $\mathbf{U} = \mathbf{S}(\mathbf{X})\boldsymbol{\epsilon}$, where $\mathbf{S} = \text{diag}(s_1, \ldots, s_p)$. Then $f_{\mathbf{U}|\mathbf{X}} = f_{\mathbf{U}|\mathbf{S}(\mathbf{X})}$. We have $f_{\mathbf{U}|\mathbf{S}}(\mathbf{U}) = |\mathbf{S}|^{-1} f_{\boldsymbol{\epsilon}}(\mathbf{S}^{-1}\mathbf{U})$. Using Lemma 12, the proof of Lemma 9 then follows along the lines of the proof of Lemma 5.